EDITED BY

SANNA OJANPERÄ,
EDUARDO LÓPEZ,
& MARK GRAHAM

OXFORD

HUMAN DEVELOPMENT AND THE DATA REVOLUTION

Human Development and the Data Revolution

Human Development and the Data Revolution

Edited by

Sanna Ojanperä
Eduardo López
Mark Graham

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© the several contributors 2025

The moral rights of the authors have been asserted.

This is an open access publication, available online and distributed under the
terms of a Creative Commons Attribution-Non Commercial-No Derivatives 4.0
International licence (CC BY-NC-ND 4.0), a copy of which is available at
https://creativecommons.org/licenses/by-nc-nd/4.0/.
Subject to this licence, all rights are reserved.

Enquiries concerning reproduction outside the scope of this licence should be sent
to the Rights Department, Oxford University Press, at the address above.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2024951954

ISBN 9780198879145

DOI: 10.1093/oso/9780198879145.001.0001

Printed and bound in the UK by
Clays Ltd, Elcograf S.p.A.

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

Contents

List of Figures vii
List of Tables viii
List of Editors ix
List of Contributors x

Introduction 1

Sanna Ojanperä, Eduardo López, and Mark Graham

PART I. THE POTENTIAL OF BIG DATA ANALYSIS FOR HUMAN DEVELOPMENT: INCREASING DIGITAL DIVIDENDS OR DEEPENING DIGITAL DIVIDES

1. **Opportunities and Challenges of Data-driven Development** 29

   Sanna Ojanperä, Eduardo López, and Mark Graham

2. **Big Mobile Data for Social Good: Opportunities and Challenges** 51

   Nuria Oliver

3. **Digital Forms of Exclusion, Big Data, and Measurement of UN Sustainable Development Goals in Africa: An Ecological Systems Theory Perspective** 71

   Bitange Ndemo

4. **Making Human Measurement Matter for Human Development: Towards a Theory of Change for the Data Revolution and "Human AI"** 90

   Emmanuel Letouzé

PART II. THE EMERGING USES OF BIG DATA ANALYSIS IN HUMAN DEVELOPMENT: CASE STUDIES FROM AROUND THE WORLD

5. **Predicting Socioeconomic Status Using Big Data** 117

   Christopher Njuguna and Patrick McSharry

6. **Reinventing the Social Contract in an Age of Big Data** 136

   Bruno Lepri, Nuria Oliver, and Alex Pentland

7. **Leveraging Big Data for Public Purposes in the Global South: LIRNEasia's Experiences** 146

Sriganesh Lokanathan, Thavisha Perera-Gomez, and Shazna Zuhyle

8. **Advancing Human Development with Artificial Intelligence: Experimental Work from Uganda** 158

Paula Hidalgo-Sanchis

9. **Extracting Insights in and around Government: Experiences in South Africa** 172

Vukosi Marivate and Nyalleng Moorosi

PART III. STRATEGIES, TOOLS, AND METHODS FOR INCLUSIVE BIG DATA ANALYSIS IN HUMAN DEVELOPMENT

10. **Digital Data Handling: A Tutorial Approach** 187

Eduardo López and Sanna Ojanperä

11. **Sources of Big Data in Global Development** 205

Eduardo López and Sanna Ojanperä

12. **Methods for the Analysis of Big Data in Development** 226

Eduardo López and Sanna Ojanperä

PART IV. ISSUES OF ETHICS, JUSTICE, AND LEGAL FRAMEWORKS RELATED TO BIG DATA ANALYSIS IN HUMAN DEVELOPMENT

13. **Data Justice for International Development: A Structural Model and Manifesto** 253

Richard Heeks

14. **The Ethics of Big Data as a Public Good: Which Public? Whose Good?** 270

Linnet Taylor

Index 288

List of Figures

2.1 Map of the impact of the floods in 2009 in Tabasco, Mexico 57
2.2 Visualization example of the SmartSteps product 61
2.3 Ground truth of crime hotspots and predicted crime hotspots 64
3.1 Aerial view of the Kibera slum before digitization 82
3.2 Kibera after digitization 84
4.1 The 4 Cs of Big Data, AI, and of the Data Revolution 95
4.2 Scatter plot of SCI vs HDI for 140 countries by region, year 2015 100
4.3 An illustration of the cognitive bias, termed the Dunning-Kruger Effect 103
4.4 The "Learning Loops" theory of change model 104
4.5 An example of generally logarithmic progress: average real GDP per capita across regions, 1870–2016 105
4.6 The three functions of data 106
5.1 Rwanda sector map showing cell tower locations 124
5.2 Scatter plots of the MPI vs the four predictive features showing their correlations 127
5.3 Scatter plot of the predicted MPI vs the actual MPI 129
5.4 Maps showing the actual (a) and estimated (b) MPI in deciles 130
7.1 Combining mobile network big data with other data can help develop rich, timely insights that serve private as well as public purposes 148
8.1 Transcription from audio into text by Pulse Lab Kampala 161
8.2 Converting unstructured audio data into structured data by Pulse Lab Kampala 163
8.3 Filtering process of public radio content with AI 165
10.1 General schematic of the data handling life cycle 191
13.1 The information value chain 255
13.2 Foundation model of structural data justice 258
13.3 First iterative development of structural data justice model 260
13.4 Second iterative development of structural data justice model 263
13.5 Revised structural data justice model 264
13.6 Forms of data justice within the revised structural data justice model 265
13.7 Incorporating explicit agency and capabilities into the structural data justice model 267
14.1 Network of participants in Responsible Data discussions, 2014–2016 274

List of Tables

2.1 Performance comparison of different models 63
4.1 Regression results for SCI and HDI on 2015 100
5.1 Censuses and surveys in Rwanda as of December 2015 118
5.2 Chronological sample of related studies and a description of the sources of data 121
5.3 Statistics from the Rwanda CDR Dataset: July 2014–March 2015 123
5.4 Summary statistics of the features and MPI 128
5.5 Estimated parameter values of the multivariate linear model consisting of four predictive features 128
5.6 Correlations of the four predictive features and the final model prediction with the MPI 129
10.1 A list of possible legal requirements necessary to handle data 194
10.2 Matrix to record the needs of a data research project 198
10.3 Key acquisition aspects of study of mobile money 200
10.4 Key data management aspects of study of mobile money 201
10.5 Aspects related to managing the end of life of data 203
11.1 Examples of Research Relating to the Sustainable Development Goals 208
12.1 Big Data methods used to address the Sustainable Development Goals 236
14.1 Interviewees contributing to the research 284

List of Editors

Sanna Ojanperä is a researcher working on her doctoral thesis at the Oxford Internet Institute, University of Oxford. Her doctoral research is funded by The Alan Turing Institute, the UK's national institute for data science and AI. Sanna is an Associate Fellow at the Digital Futures at Work Research Centre (Digit). She also consults as a researcher and a data scientist for development organizations.

Eduardo López is an Associate Professor in the Department of Computational and Data Sciences at George Mason University. He studies collective phenomena in natural and socioeconomic systems, patterns that emerge out of the interaction of collections of similar entities (people, institutions, or particles). In his work, Prof. López employs different methods including mathematical and computational modelling, as well as large-scale data analysis.

Mark Graham is the Director of Fairwork and Professor of Internet Geography at the Oxford Internet Institute, University of Oxford. Mark is an economic geographer with an interest in how digital technologies and digitally-mediated practices affect economic and social inequalities. His research focuses on economic development, labor, power, participation, and representation.

List of Contributors

Richard Heeks is Chair in Digital Development at the Global Development Institute, University of Manchester; and Director of the Centre for Digital Development. He has been consulting and researching on digital and development for more than 30 years. His most-recent book publication is "Information and Communication Technology for Development".

Paula Hidalgo-Sanchis has worked over 20 years integrating innovation into humanitarian and sustainable development practice in Africa, Asia, Latin America, and The Caribbean. She is passionate about promoting the responsible use of innovations and Artificial Intelligence-based technologies for sustainable development, humanitarian action, and peace.

Bruno Lepri is a senior research scientist at Fondazione Bruno Kessler, where he leads a group devoted to computational social science, human-centric machine learning, and cooperative AI. He is also advisor and consultant of several companies and organisations on AI topics.

Emmanuel Letouzé is the Director and Co-Founder of Data-Pop Alliance. He also co-founded the Open Algorithms (OPAL) project and served as the Executive Director of the OPAL pilots in Senegal and Colombia in 2017-2020. His work focuses on the applications and implications of data, statistics, technology and AI for development and democracy in the Global South.

Sriganesh Lokanathan is a researcher, practitioner, and interlocutor at the nexus of innovation and global development. He co-established and led LIRNEasia's big data for development work. He conceptualizes and implements responsible digital innovation to support global development whilst understanding its implications on society and mainstreaming its use for social good.

Vukosi Marivate is an Associate Professor of Computer Science and holds the ABSA UP Chair of Data Science at the University of Pretoria. He specialises in developing Machine Learning (ML) and Artificial Intelligence (AI) methods to extract insights from data, with a particular focus on the intersection of ML/AI and Natural Language Processing (NLP).

Patrick McSharry is a Professor at Carnegie Mellon University. His research focuses on artificial intelligence, machine learning, data science, big data, forecasting, predictive analytics, and the analysis of human behavior. He is currently building multidisciplinary partnerships to create data-driven and sustainable solutions to global challenges.

Nyalleng Moorosi is a senior researcher at Distributed Artificial Intelligence Research (DAIR) institute. Her interests are in understanding how we can build models which center populations often regarded as peripheral. Before DAIR she was a research software engineer at Google, where she was one of the first employees at the Google Africa research lab.

Bitange Ndemo is Kenya's Ambassador to Belgium and the EU. He was Professor of Entrepreneurship at the University of Nairobi and chaired Kenya Distributed Ledgers and Artificial Intelligence Taskforce. He is a member of the MIT Artificial Intelligence Policy Forum. Formerly he served as Permanent Secretary in Kenya's Ministry of Information and Communication.

Christopher Njuguna is an information technology practitioner and data scientist with a passion for developing and building solutions to Africa's challenges. He currently leads the engineering and data teams at Qhala Ltd.

Nuria Oliver is the Co-founder and Director of the "Institute of Humanity-centric AI" (ELLIS Alicante Foundation) and holds a PhD in Artificial Intelligence from MIT. She is an IEEE, ACM, EurAI, and ELLIS Fellow and an elect member of the Royal Spanish Academy of Engineering and SIGCHI Academy. She is the author of 180 publications, the book "Artificial Intelligence, naturally", and the inventor of 41 patents.

Alex Pentland is professor at MIT, former Academic Head of MIT Media Lab, and one of the most cited authors in computational science. He is advisory board member for UN Global Partnership for Sustainable Development Data, and formerly for American Bar Association, Google, AT&T, etc., and is member of U.S. Academy of Engineering.

Thavisha Perera-Gomez was a research manager at LIRNEasia where she explored the applications of new data sources, specifically big data, for achieving the Sustainable Development Goals. She has also worked in strategy consulting and conducting industry research. She holds an MBA from the Postgraduate Institute of Management, Sri Lanka.

Linnet Taylor is Professor of International Data Governance at the Tilburg Institute for Law, Technology, and Society (TILT), in the Netherlands. Her research focuses on digital data, representation and democracy, with particular attention to transnational governance issues.

Shazna Zuhyle has 15+ years' experience managing projects in the ICT domain. At LIRNEasia she led the research on broadband QoS for fixed and mobile broadband services in Asia and the Pacific, and managed the data, analytics, and policy team. She holds an MSc in Telecommunications from University College London.

Introduction

Sanna Ojanperä, Eduardo López, and Mark Graham

We live in a world where the ability to collect and use data about humans and their environment on a large and ever-growing scale has revolutionized the way societies, economies, and politics function and develop. While keeping records and storing information is characteristic of every society, the role of data has grown exponentially through the last two phases of industrial change. Starting in the latter half of the twentieth century, the Digital Revolution introduced computers and other digital devices and connected the world via the Internet. It recast economies and societies through an increased focus on information technologies thus heralding the beginning of the Information Age. At the start of the twenty-first century, the so-called Fourth Industrial Revolution broadened these changes through increased automation and deeper interconnectedness which have resulted in ever more rapid technological change. The emergence of new technologies such as artificial intelligence (AI) and advances in biochemistry and robotics soften the boundaries between the physical, biological, and the digital, and have led to seismic shifts to economic and societal patterns. With two-thirds of humanity connected to the Internet, many around the world now lead lives where connectivity is constant and embedded to a significant part in their economic, social, and political realities.

While the two phases of industrial change are defined by distinctive characteristics, they are both underpinned by the ability to record, store, and utilize ever increasing amounts of data. Data can describe phenomena and their analyses enables intelligent decision-making and construction of processes that are founded on evidence. Accurate and true data and their appropriate use are particularly valuable at a time when the abundance of communication technologies and data-generating processes results in vast volumes of data and have given rise to misinformation and disinformation. While the volume of available data is unprecedented, there are significant gaps in terms of what phenomena are covered (Kitchin, 2014), how geographically and demographically even that coverage is (Graham et al., 2019; Graham & Dittus,

Sanna Ojanperä, Eduardo López, and Mark Graham, *Introduction*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Sanna Ojanperä, Eduardo López, and Mark Graham (2025). DOI: 10.1093/oso/9780198879145.003.0001

2022), and how those data are used (United Nations, 2023). The Digital Revolution and the Fourth Industrial Revolution have not had a uniform impact around the world but have arrived with different speeds and varied in their influence. Significant parts of the world remain disconnected, and risk not being reflected in the data that are used to power processes and inform decision-making. Furthermore, where sufficient capacities are not developed to critically evaluate data-driven processes, the risks and harms related to improper use of data present real dangers to the economy and society. Understanding and addressing these digital divides and reaping the benefits of the data revolution is high on the policy agenda of many countries around the world who seek to foster an inclusive economy and society in the "age of data."

There are few areas where the question of the reach of the data revolution is more relevant than in global development.[1] Concerned with understanding and addressing the uneven levels of economic and human development within and between countries, those working within this field would benefit from more meaningful data to use as evidence when designing projects and policies as well as for monitoring and evaluating their implementation.

Has there been a data revolution in global development? The relevant actors involved are diverse, spanning many institutional levels such as local staff working in development projects; national statistical agencies and policy makers in respective governments; development partners and development cooperation producers working bilaterally or multilaterally; international agencies and organizations working with a global view; and academia, civil society, and the private sector researching related topics whether within or outside the country. The capacities to collect and use data differ among these actors, and as smaller scale data and traditional statistical methods to analyze them form a touchstone of much of the work carried out within global development, many actors are not used to handling large datasets. As the value of data is increasingly recognized within the field, particularly over the past

[1] Many actors working with development use terms such as international development and global development interchangeably to refer to the concept of uneven levels of economic and human development around the world, the development challenges underpinning those levels, and to attribute the industry working to address these challenges and the uneven levels of development as well as the associated field of research and academic discipline. Some working on these topics advocate for the use of the term global development, as within-country inequalities have grown to rival between-country inequalities in many areas of development and as the term better captures the view that sees development challenges such as environmental sustainability, poverty, economic security, and migration as universal issues affecting countries around the world (Horner & Hulme, 2019). One could see the adoption of the SDGs around the world as an example of a global view of development. We acknowledge that choosing accurate and appropriate terms to refer to these complex topics is challenging but have decided to use the term global development throughout the chapters we have authored in this book. While developing this edited volume, we found that many of the topics pertaining to data driven development concern countries around the world and across all income groups. However, we would like to urge readers to make themselves familiar with the contested nature of these terms and join us in actively seeking for suitable substitutes.

decade or two, elements of analysis of large datasets have begun to be reflected upon or integrated in the varied roles related to development work as well as in the associated training pathways.

The development community took a particularly big leap towards data when adopting the 2030 Agenda for Sustainable Development in 2015. The 2030 Agenda sets global development priorities whose achievement requires a significant effort in measuring development outcomes across the 17 goals. Relevant data are required to identify needs and prioritize between the goals, monitor progress, facilitate decision-making, and promote discussion and dialogue. Such relevant data could come from official sources as well as from the private sector or civil society and through many online sources. What is certain is that in order to achieve the Sustainable Development Goals (SDGs) or to expand skills in handling and utilizing large data within global development, technical training is required to build capacity to understand and utilize various data sources. In a report published in 2017, the OECD highlighted that development cooperation must take advantage of the data revolution by applying new datasets, expanding the traditional methodological toolkit to include new methods to produce and use statistics, and forge partnerships with other communities of producers and users of data (OECD, 2017). This book is intended to help those working with global development or who are interested in the area to familiarize themselves with big data,[2] its applications, suitable methods, and the other actors in the data ecosystem who could make potential partners for those working with global development. While the rest of the world embarks on adopting data-driven or data-intensive approaches to their work, actors within global development should ensure that they too unlock the value of data and do not miss out on crucial insights and evidence that big data can offer to address pressing development challenges. The volume of available data is not about to slow down, and joining the processes and discussions to shape their use towards more equitable ways that benefit a wide set of people should be high on the agenda of all development actors.

The Focus of the Book

"Human Development and the Data Revolution" tackles a topic that in the age of data and digitalization has become timely in global development: what are the potential uses of large-scale data in the contexts of

[2] Please see the following section for a definition of "big data."

development, in particular, what techniques, data sources, and possibilities exist for harnessing large datasets and new online data to address persistent concerns regarding human development, inequality, exclusion, and participation. The volume also sheds light into the challenges and obstacles related to the safe and effective use of big data and computational approaches in development, given that safeguarding privacy and security as well as ensuring accurate insights are crucial principles when addressing development issues that often concern vulnerable or underserved populations.

The book directly addresses these questions, employing a global perspective to explore the latest advances at the intersection of big data analysis and human development. It is unique in combining research perspectives on the potentials and limitations of big data analysis for human development, as well as practical contributions on methods and case studies, making it accessible to development researchers, scholars, practitioners, policymakers, and anyone interested in the topic.

Bringing together perspectives from academia and development practice, the book has a two-pronged focus on theoretical and practical research on big data and computational approaches in human development. It details the potentials of big data analysis for development and discusses how the challenges or barriers to the adoption of big data analysis might present limitations for its usefulness in contexts of development. The book reviews a range of methods and tools that have been employed to increase local ownership of big data and presents research results of projects employing big data in contexts of development. The book's combination of theoretical and practical findings at the intersection of big data analysis and human development is a unique crossover and makes for a singular contribution with considerable appeal across the disciplines of global development, geography, and informatics. Various books have been published on the technicalities of carrying out big data analysis and which cater to various audiences from beginners to more advanced professionals. Separate sets of instructive books have been published on particular technical areas of interest such as programming languages popular in data science and computational research, Apache Hadoop distributed data infrastructure, and Apache Spark cluster-computing framework. Another related set of recently published books addresses big data analysis and computational research on a general conceptual level and are targeted at the general audience. Despite the increasing demand for big data analysis in the context of development theory, practice, and policy, none of these works relate to the domain of global development. This volume complements these books by catering to this field and by meeting the requirements of both non-specialist audiences as well as those with advanced knowledge of

big data analysis and computational approaches. What sets the volume apart is its unique character of being firmly embedded in the context of theoretical and empirical development research as well as practical operational work and development policy.

In the next sections of this introductory chapter, we will define the key concepts of big data, development, and data-driven development in more detail, present the analytical focus and the conceptual framework of the book, introduce the four parts of the book and the constituent chapters making up each part, discuss what the reader should expect from reading this book, and review the goal of what we hope to achieve by bringing these perspectives together inside one cover.

Defining Key Concepts about Big Data and Global Development

The Advent of Large Complex Data and of Suitable Analytical Approaches

The increased availability of data over the past couple of decades has been driven by several technological changes. These developments concern great leaps in both hardware (the physical components of a computer such as the central processing unit that allows data to be processed or the screen that lets a user communicate with a computer) as well as software (the programs and applications that run on the hardware such as the operating system, or statistical programming suites such as Stata, R, or Python). Some of the factors underpinning the speed of this change are the approximate doubling of the performance of computer chips approximately every two years (also called Moore's law) leading to cheaper and faster computing and the wider availability of interconnected devices to collect data such as sensors and the many appliances pertaining to the "Internet of things." With the rapid evolution of hardware components and the software to exploit their increased capacities, there have been marked increases in the collection and flow of data, the storage capacities to hold that data, and the processing power and facilities to use it. The digitalization of many processes across the public and private spheres has also enabled the rise of platforms where data are generated through digitally enabled interactions, transactions, behaviors, and production (such as social media or remote work platforms) or reported by users as citizen science (such as platforms recording environmental observations). Additionally,

many previously non-digitally stored data sources are being made available through digitization, and as historic records may stretch out far into the past, this digitized historic data further increase the volume of available data.

The growing volume of these data led to the adoption of the term "big data" almost thirty years ago.[3] There are many competing definitions for big data, but most of them tend to refer to a number of "Vs" characterizing the concept,[4] including volume (emphasizing the large size of the datasets, often measured in terabytes or petabytes), velocity (highlighting the fact that many sources produce data continually, for instance through sensor recordings, as well as the high frequency at which data are handled or made public), variety (data derive from diverse sources and may range between structured, semi-structured, or unstructured types), veracity (referring to the reliability and quality of the data), and value (representing the informational and monetary value derived from the analysis of large datasets). For the purposes of this book, we will be referring to these large datasets as both "big data" and "large-scale data." In order to allow the examination of a wide variety of relevant contexts and examples we will define the terms fairly loosely in line with the findings of De Mauro et al. (2015) as consisting of data of a size that exceeds the capabilities of most traditional statistical software suites that are commonly used by development professionals such as Excel, SAS, SPSS, or Stata, and which can inform research questions about human well-being regardless of the source from which they are derived.

The increased availability of data combined with the enhanced abilities to store and process them have led to a notable change in the attitudes towards decision-making. In fact, while a decade ago "big data" may have been the buzzword in private and policy circles, the crux of the "big data paradigm" is not the large amount of data itself, but the prospects of analyzing them to derive evidence to benefit decision-making. The key feature of this paradigmatic change is therefore that the use of these large datasets is systematically applied to inform intelligent decision-making (Hilbert, 2016). Many traditional statistical tools fail to derive insights from large and often complex datasets, and specific analytical methods are required. The advent of large

[3] The first use of the term "big data" in ACM journals dates back some 30 years ago to an article by Cox and Allsworth (1997) but the term was popularized over the following decade within policy and private circles. However, the underlying phenomenon—advances in technological development enabling increasing volumes of data which in turn are used to inform decision-making—has been described by other names much earlier such as "information explosion" which first appears in Oxford Dictionary in 1941 (OED Online, 2022).

[4] The "Vs" related to big data often refer to three, four, five, or ten characteristics beginning with V, but as big data continues to be researched, authors have discussed lists reaching up to 56 characteristics (Hussein, 2020).

datasets has therefore introduced the need for analytical capabilities that combine skills from computer science with analytically oriented approaches ranging from business studies to social research. These analytical capabilities are often termed data science, data engineering computational social science, or computational analysis. The proliferation of both large complex datasets and the rise of data science has in turn facilitated the development of other technologies such as algorithmic approaches to data management and analysis, in particular tools and approaches rooted in AI. In order to make a distinction between the data sources and the wider development, Letouzé has suggested the adoption of the term "big data" (in lower case) to refer to the large scale data sources specifically and the term "Big Data" (capitalized) to refer to the changing ecosystem, with four constitutive elements: (1) the "crumbs" of big data acting as a fuel; (2) the surrounding evolving infrastructural and human capacities; (3) communities comprised of contributors, users, and developers of data-driven analytical systems; and (4) the culture constituting the incentives, expectations, and norms that arise from and shape the use of data-driven systems (see Chapter 4 for a further discussion on the topic).

In fact, with the burgeoning adoption of large-scale data as part of many organizations' operations, the discourse in policy and private circles has perhaps widened beyond "big data" per se, and rather many frontiers of work now utilize terms such as data science, AI, algorithmic decision-making, or computational research. In such work, a focus on large datasets is often implied, as the methodologies associated with these topics were largely created to manage big datasets. While we do discuss big data analysis and refer to particular analytical approaches such as data science or computational research within this volume, one should note that these are very much evolving terms. Some novel methodologies to manage large datasets have been developed relatively recently when the volume of available data began to surpass the capacities of popular statistical software packages. While there are many established analytical approaches and methods, as discussed in more detail in Chapter 12, there is also an active research frontier in methodological development related to these topics. In order to reflect this evolution and to facilitate the examination of a wide variety of relevant contexts and examples where disciplinary and contextual traditions may prefer certain terminology over another, we note that some sections of this book discuss data science, artificial intelligence, algorithmic decision-making, and computational analysis, while others may refer to big data analysis. We have asked each contributing author to specify the analytical approaches and tools they use in order to reach across the varying terminological

traditions. Additionally, Chapters 10, 11, and 12 contain detailed discussions on data sources, methodological approaches, and considerations related to data management in order, respectively, to facilitate the reader's grasp of the technical landscape.

While big data analysis is fast being adopted as a key feature of many entities both private and public, this is not to say that doing so is effortless or straightforward. On the contrary, the fact that technical and domain-specific expertise and deep skills are required to derive benefit and avoid shortcomings, gives rise to various challenges and barriers that complicate the integration of big data and computational approaches. While many barriers impeding the adoption of big data analysis can be identified, five general groups of barriers have been proposed, including technical (such as those related to the extent of infrastructure capabilities to manage and process data); economic (such as financial resources, associated costs, and return on investment); organizational (such as those related to organizational structure, culture, political buy-in, and organizational policies including those related to security, privacy, data governance, and ownership); external context (such as the policy and legal landscape, public IT infrastructure, and the supply of service providers and vendors specialized in big data); and people-related barriers (such as public awareness, skills, experience, and training) (Alalawneh & Alkhatib, 2021; Al-Qirim et al., 2017; Kankanhalli et al., 2016; Sagiroglu & Sinanc, 2013). Focused on low and middle income countries,[5] Alalawneh and Alkhatib (2021) found that the relative importance of the barriers changes between the contexts under study. The authors report that public sector entities seeking to adopt big data analysis found the organizational and people-related barriers to be most prominent of the five types, while barriers related to external context were identified as most significant among service sector firms and economic barriers were found to be prominent across all sectors.

Big data are not a silver bullet and using big data sets to extract insights for human development presupposes an ability to organize and manage large and potentially complex datasets within the frameworks of social research approaches (Einav & Levin, 2014). Further, given that big data are often

[5] We use the term low-and middle-income countries (LMICs) to refer to the countries that fall into the economic classification of low-income, lower-middle-income, and upper-middle-income as defined by the World Bank (2023). While many actors working in development will employ terms such as Global South (as opposed to Global North), developing countries (as opposed to developed countries), or Southern countries (as opposed to Western countries), we are aware of the problematic nature of these terms and their associated dichotomies, which is why we chose to not use them in the chapters we have authored in this book. However, we acknowledge that selecting accurate and appropriate terms is challenging and we would like to urge readers to make themselves familiar with the problematic nature of the terms we have decided to not use and join us in actively seeking for suitable substitutes. For anyone interested in delving into further perspectives on this topic, we recommend the following pieces from Kloß (2017), Haug (2021), and Sud & Sánchez-Ancochea (2022).

collected for a purpose different than the uses they may find in global development, they may not offer "a perfect natural experiment" to help answer all relevant causal questions (Imbens & Angrist, 1994) that development actors may seek to address. However, even where the true underlying dynamics of a particular research question may not be addressable with the analysis of certain big datasets, the insights that can be derived may offer much useful peripheral information that can shed light to better understand complex development challenges.

Global Development and Large Complex Data

Global development is concerned with enhancing human and economic development, reducing poverty, and increasing well-being and prosperity. Development work tends to take place through addressing issues that have been identified within areas such as education, health, and food security through specially designed projects and programs targeting the most deserving beneficiaries. Owing to the issue-focused and program- and project-based nature of much of development work, the field pays inherent attention to data and measurement. Statistics produced by national statistical institutions and international organizations as well as surveys administered to the target demographics of development interventions account for large parts of the traditional data produced and analyzed within this field.

However, the advent of data and new analytical methods has not passed the field by. Indeed, parallel to the wider waves of technological change outlined above, new ways to approach development work have emerged in the community of practice. Popularized through their acronyms or monikers, the field has seen a succession of themes surfacing at the intersection of various technological trends related to data availability, such as: IT4D focusing on information technology/ies for development; ICT4D focusing on information and communication technology/ies for development, see e.g., (Heeks, 2017; Unwin & Unwin, 2017); digital development focusing on digitalization and digital tools and their connection to development, see e.g., (World Bank, 2016); big data 4 development (BD4D) and data 4 development (D4D) focusing on large datasets for development, see e.g., (Kshetri, 2014; Hilbert, 2016), AI4D focusing on the application of AI for development, see e.g., (Mann & Hilbert, 2020), ML4D focusing on the application of machine learning for development see e.g., (Esteban et al., 2022), IoT4D focusing on the Internet of Things for Development, see e.g., (López-Vargas et al., 2020) as well as related concepts data 4 good, data 4 social good, tech 4 social good (T4SG), data science for social good (DS4SG), and AI 4 social good

(AI4SG), see e.g., (Tomašev et al., 2020; Cowls et al., 2021), which address the application of data or technology solutions to aid with achieving a particular social goal.

While these themes vary between the constellation of involved actors; the range of dedicated events, conferences, or fora; the extent of publication and outreach on the topic; and the amount of earmarked funding, there is often a focus on developing, applying, and educating others about relevant tools and methods through applications and case studies. In terms of their analytical foci, they tend to investigate and address digitalization and adoption of communication technologies and digital devices, paying particular attention to the extent to which digital divides exist between those who can connect and do participate and others who do not. While many of these themes tend to share an optimism about the value of data and often advocate for the greater use of data in understanding and addressing particular development issues, they are also taking note of critical interventions focusing on the harms and risks of data (Dalton & Thatcher, 2014; O'Neil, 2016; Noble, 2018; Benjamin, 2019; D'ignazio & Klein, 2020; Graham & Dittus, 2022).

In recent years, the value of large-scale data for development has been recognized by many key development partners. Following the establishment of the United Nations Global Pulse as an initiative focused on leveraging big data and AI for development in 2009, the United Nations Secretary-General appointed an Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG) (IEAG, 2014) in 2014. The UN Global Working Group on Big Data for Official Statistics was established in the same year to oversee how harmonized statistical frameworks may benefit from the introduction of new statistical products and methods utilizing big data (UN-CEBD, 2023). The multistakeholder network Global Partnership for Sustainable Development Data was launched in the following year in 2015 to help support the use of data and facilitate the creation of data partnerships (GPSDD, 2022). While explicitly focusing on digital development, the OECD's Going Digital Integrated Policy Framework (OECD, 2021) and the Principles for Digital Development conceptualized by a dedicated working group (Digital Impact Alliance, 2023) both highlight the value of large-scale data analysis. Many development cooperation partners have similarly adopted frameworks and guidelines to help introduce digitalization and data-driven approaches to their work.

While there is significant interest in adopting big data and data science methods within global development, the experience and information are

shared in disparate venues within academia, policy circles, and the private sector, complicating knowledge transfer and making reflection on the overall developments within the field challenging. Within academic work, there are no books considering the topic holistically from a variety of sectoral and institutional viewpoints. The available books tend to focus on the application of big data analysis in certain sectors such as urban sustainability (Wang, 2018), sustainability analysis (Dunn & Balaprakash, 2021) environmental analysis (Uitto, 2021), agriculture, health, and environment (Kshetri, 2016), policy and democracy (Bertoni et al., 2023), social science (Jemielniak, 2020), primary industries, education and R&D, and e-commerce (Kshetri, 2017), or discuss big data and data science in relation to measurement challenges such as data collection in fragile countries (Hoogeveen & Pape, 2019) or building statistical capacity (Thurston, 2020; Vance et al., 2022). The academic articles focused on big data for development mainly concentrate on clarifying concepts, developing theoretical frameworks, addressing the challenges and opportunities presented by big data analysis (Kshetri, 2014; Ali et al., 2016; Arora, 2016; Hilbert, 2016), or investigate the value or applications of big data and data science for specific sectors or topics. In Chapters 11 and 12 of this book, we refer to numerous examples and case studies often published as academic articles, and which span across data sources, methods, and touch on topics framed by each of the 17 SDGs.

The body of knowledge addressing big data and data science in development rests only partially within academic publishing. In particular, reports combining policy advice and research have been produced by international organizations, with IEAG's report entitled "A World That Counts" (United Nations, 2015), and the UN Global Pulse's reports on big data (2012, 2013) having opened many discussions a decade ago. A few years later, the UN Development Group (2017) published a document offering guidelines on data privacy, data protection, and data ethics in the use of big data deriving from private sector entities, which was followed by the UN's (2018) adoption of the Principles on Personal Data Protection and Privacy and the UN Global Pulse's (2018) development of similar principles for their work. The topic has continued to receive attention also more recently in some of the flagship publications read by those working in development cooperation, such as the Development Co-operation Report 2017: Data for Development (OECD, 2017) and the World Development Report 2021: Data for Better Lives (World Bank, 2021). The inclusion of many examples of projects applying big data analysis within these reports illustrates the diversity of development work applying these approaches and the many development partners interested in them.

Supporting knowledge transfer between the fractionalized field of global development, conferences and events focused on big data and data science in development bring together actors from academia, public and private sectors, civil society, as well as various development partners. Organized by the UN's Statistical Division, and directed to those working with official statistics, the International Conference on Big Data and Data Science for Official Statistics has met annually since 2014 and the United Nations World Data Forum, directed to a wider audience, has met biannually since its inception in South Africa in 2017. These events have been seen as catalyzing data innovation and building political and financial support for data. The World Data Forum also launched the Cape Town Global Action Plan for Sustainable Development Data (HLG-PCCB, 2017), which is intended to provide a framework to support statistical capacity building in six strategic areas to achieve the 2030 Agenda. Interesting, related work is also carried out by nonprofits such as DataKind, which facilitates organizations' use of data science to address humanitarian challenges (2023), and Flowminder, which focuses on big mobile data to improve health and welfare in low- and middle-income countries (2023).

The Book's Analytical Focus and Framework

The analytical focus of this volume seeks to better understand the nexus where three central themes connect: first, the field of big data analysis is continuously progressing and enjoys growing popularity around the world; second, the development community's interest in integrating data-driven approaches to their work is increasing; and third, concerns related to the safe and ethical use of big data are emerging. This volume is further underpinned by a framework that seeks to investigate central topics that on one hand give optimism for potential impacts of big data to be reaped by development stakeholders and on the other highlight some of the barriers and challenges that a development actor seeking to introduce big data analysis should be aware of and prepared to navigate. The framework comprises the following themes:

Potentials of Big Data Analysis for Development

The book presents a range of case studies and examples that discuss the central avenues through which big data can add value to development work, focusing on three principal areas:

(1) Enabling operational and strategic insight and improving quality of development outcomes through:
   a. generating new understanding on an existing topic by adding finer detail into a previously coarse level of analysis;
   b. allowing the investigation of completely new topics identified through previously unused data sources;
   c. increasing the geographical coverage and granularity of development data.
(2) Enabling the analysis of real-time or near real-time events given the more frequent recording of observations characteristic of many big data sources.
(3) Enabling cost-savings through exploiting the nature of big data and the automatability of analytical tasks.

Challenges and Barriers of Big Data for Development

The chapters also address five central challenges relating to data-driven development and representing the five main categories of barriers associated with the adoption of big data:

(1) Representativeness and quality of the big data sources (barrier related to data);
(2) Differing capabilities of development stakeholders to produce and consume big data analysis (barrier related to people);
(3) Securing political commitment (organizational barrier);
(4) Financial resources required to integrate big data analysis (economic barrier);
(5) Regulatory and ethical aspects of big data in development (barrier related to external context).

The themes comprising the book's framework are discussed in more detail in Chapter 1. While other opportunities (such as new products and business models) and challenges (such as surveillance, exploitation, and ownership or expropriation) related to big data can be conceptualized, we have included these themes because they were the central topics that our discussions with the authors and our desk research brought up. The chapter also includes a brief reflection describing how the book's remaining chapters address the themes of the framework and what are some of the central insights they offer. Many chapters touch on various elements within the framework, and we have therefore structured the book not according to the framework, but according

to a four-part structure based on the kind of contribution the chapters offer. The following section presents the structure of the book and introduces the individual chapters and their authors.

Parts and Chapters Comprising This Book

Beyond this introductory chapter, this book consists of four distinctive sections: Part I focuses on the potentials of big data analysis for human development; Part II describes the emerging uses of big data analysis in human development through case studies from around the world that reflect on both opportunities and challenges of integrating large data into development; Part III outlines the strategies, tools, and methods for inclusive big data analysis in human development; and Part IV concludes with a discussion on the themes of ethics, justice, and legal frameworks related to big data analysis in human development.

Part I: *The Potential of Big Data Analysis for Human Development: Increasing Digital Dividends or Deepening Digital Divides* has a twofold focus: it describes potentials of big data analysis for human development and situates that discussion in the current data landscape in terms of technological development, infrastructure, and human capital. This section features the following chapters:

In Chapter 1, Sanna Ojanperä, Eduardo López, and Mark Graham, the editors of the volume and researchers affiliated to the University of Oxford (Sanna and Mark) and George Mason University (Eduardo) examine the themes comprising the conceptual framework underpinning the book in more detail. They reflect on the key themes of the book through the viewpoints offered by its 14 chapters and draw together central findings. In Chapter 2, Nuria Oliver, the Co-founder and Director of the Institute of Humanity-centric AI at the ELLIS Alicante Foundation discusses the ways in which big mobile data might facilitate better decision-making and positive social impacts. She also highlights a diverse set of challenges that need to be addressed to realize the full benefits of the data captured by mobile network infrastructure. Her chapter offers an account of two projects using big mobile data for natural disaster management in Mexico and crime detection in the United Kingdom and discusses the opportunities and challenges from a methodological perspective. In Chapter 3, Kenya's Ambassador to Belgium and the EU and the former Permanent Secretary in Kenya's Ministry of Information and Communication, Bitange Ndemo, investigates leveraging big and open data to achieve SDGs. He makes a case for increased use of

big data in African decision-making and shares a policy-maker's perspective into the sometimes challenging process of securing political will. Drawing from his extensive career in the Kenyan government, Ndemo illustrates how African policymakers approach the collection and integration of big data in decision-making and describes the political economy of measurement choices. Concluding Part I of the book, in Chapter 4, Emmanuel Letouzé, the Director and Co-founder of Data-Pop Alliance dives into the history of global development data and related measurement and evaluation efforts. Mirroring that history against current uses of big data and AI, he proposes theoretical concepts to enhance the role of data and the use of empirical evidence for human development.

Part II of the book, *The Emerging Uses of Big Data Analysis in Human Development: Case Studies from Around the World*, discusses research results as well as any opportunities and challenges identified. The case studies draw together examples from more academically oriented perspectives as well as from viewpoints closer to the practitioner landscape. In sharing these concrete examples, this section illustrates how approaches based on big data are adopted and what impacts they have in development contexts, where previous technologies have been used to varying extents.

Opening Part II of the book, in Chapter 5, Christopher Njuguna, an information technology practitioner and engineering and data science team lead at Qhala Ltd and Professor Patrick McSharry, a Professor at the Carnegie Mellon University Africa, provide a proof-of-concept tool for measuring multi-dimensional poverty in Rwanda through two call detail record (CDR) based features: mobile ownership per capita and call volume per phone, which they combine with satellite nightlight data and population density data. The chapter offers a focused overview of the benefits and challenges of using big data for measuring poverty and discusses how they can be used alongside household surveys in order to provide timely estimates of socioeconomic development. In Chapter 6, Bruno Lepri, Senior Research Scientist at the Fondazione Bruno Kessler, Nuria Oliver, the Co-founder and Director of the Institute of Humanity-centric AI at the ELLIS Alicante Foundation, and Alex Pentland, Professor at MIT and former Academic Head of MIT Media Lab, share a detailed description of the Open Algorithms (OPAL) project, which aims to open and leverage private sector data for purposes of public good and which has been piloted in Senegal and Kenya. The authors also discuss the issues and challenges introduced by data-driven and algorithmic approaches to policymaking in development. Sriganesh Lokanathan, a former Policy Fellow, Thavisha Perera-Gomez, a former Research Manager, and Shazna Zuhyle, a former Research Fellow at LIRNEasia present an

overview of the emergence of, work carried out by, and unique opportunities and challenges faced by the Sri Lankan think-thank in Chapter 7. LIRNEasia leverages big data for public purposes and is a leading institute carrying out deeply skilled data-driven work across Asia-Pacific. The chapter offers a unique perspective into founding and developing a think-thank focused on big data and addressing constraints related to skills, resources, and political support. In Chapter 8, Paula Hidalgo-Sanchis draws from her experience managing the United Nations Pulse Lab Kampala and discusses an AI-tool that the Lab developed in Uganda to process audio files of radio conversations into text files containing testimonies, opinions, and comments of people participating in radio programs. The research team extracted testimonies, opinions, and reports relating to three areas of interest: client feedback on public service delivery, the impact of societal behavioral change campaigns, and an early warning system for managing accountability and risk. The chapter discusses the application and the usefulness of the kinds of text files extracted and touches on the ethics involved in processing unstructured audio data into structured text files. Concluding Part II, in Chapter 9, Vukosi Marivate, Associate Professor of Computer Science at the University of Pretoria, and Nyalleng Moroosi, Senior Researcher at the Distributed Artificial Intelligence Research Institute, review the opportunities and challenges related to a selection of data science initiatives undertaken by the National Scientific and Industrial Research Council in South Africa. Topics covered include building up a pipeline for data science education and training, access to data and institutional readiness, managing relationships with government partners, and balancing between the goals and incentives of researchers and policymakers.

Part III, *Strategies, Tools, and Methods for Inclusive Big Data Analysis in Human Development* focuses on methodological issues specifically relating to big data in global development. It reviews the tools, methods, and strategies for big data analysis in development and in particular in the context of operational projects and initiatives. The three chapters discuss such themes as data acquisition, data management, data mining, and present a selection of the methods used in analyzing big data for development. These themes are discussed in light of practical applications in development projects and initiatives as well as studied in terms of their strengths and weaknesses.

In Chapter 10, Eduardo López, Professor at George Mason University and Sanna Ojanperä, Doctoral researcher at the University of Oxford and The Alan Turing Institute, offer a view into the considerations and processes underpinning safe and effective research with big development data. They discuss the data handling life cycle and illustrate how the characteristics

of the data and the intended uses and users impact it. Formal and legal considerations that pertain to data handling are also discussed. The authors additionally present a sample use case to help the reader grasp how these concepts can be applied when designing a study or a project using big development data. In the following Chapter 11, the authors discuss the notion of big data from the point of view of both data researchers and practitioners, as well as the new attitude that has taken hold with respect to the use of these large data sources. They then introduce the reader to a selection of the most popular big data sources as applied in development research and projects and illustrate them with examples drawing from all of the 17 SDGs. Concluding Part III of the book, in Chapter 12, López and Ojanperä focus on analytical approaches to extract meaning from big data in development. They relate this discussion to the skills and training landscape of development professionals. The reader is introduced to the probabilistic foundation of the analytical methods, and the evolution of data from "small" to "big" and how this has changed methodologies and ways of thinking. The authors then present a selection of analytical approaches that have been used in development work that incorporate big data and illustrate them with examples drawing from all of the 17 SDGs.

Part IV, *Issues of Ethics, Justice and Legal Frameworks Related to Big Data Analysis in Human Development*, focuses on the intersection of issues of ethics, justice, privacy, and regulation with human development—a domain that holds notoriously sensitive data about a part of the population that often faces vulnerability. The two chapters chart the development of regulatory and policy frameworks governing big data in development and interrogate how data can be conceptualized, viewed, and treated through a lens focusing on its social implications. They draw from examples from the private and public spheres and discuss the roles and perspectives of various development stakeholders and other central actors in the data ecosystem such as tech companies.

In Chapter 13, Richard Heeks, Chair in Digital Development at the Global Development Institute and Director of the Centre for Digital Development at the University of Manchester, revises one of the structural data justice models based on field studies of data-intensive development carried out in India and Kenya. He offers the revised model as a conceptual framework for research around the topics of data-intensive development. Heeks then uses the model to derive a "Data-Justice-for-Development Manifesto," which he invites to be used to guide development policy and practice. In Chapter 14, Linnet Taylor, Professor of International Data Governance at Tillburg University, evaluates how the argument for "data as a public good" fits with the corporate

experience of big data and explores existing models for data sharing. The author grounds her analysis in the idea of corporate data as an ecosystem involving rights, duties, and claims that often conflict with one another. She evaluates the power dynamics that result from conceptualization of data as a public good and analyzes the different ethical stances authors adopt in relation to its use.

What the Reader Should Expect to Gain from Reading the Book

This edited volume offers the reader three key features: (1) unique book focus and policy implications, (2) familiarization of the research frontier of big data and development, and (3) an accessible account of methodological approaches. The contributing authors hail from diverse backgrounds in academia, development practice, civil society organizations, government, and the private sector. Many of them are pioneers in building critical foundations of theoretical and practical knowledge on the emerging field of big data in development, which makes them uniquely suitable contributors to the volume. This book is distinctive in gathering many of these pioneering voices into one volume and in introducing the reader to the frontier of research and practice on the topic.

Unique Book Focus and Policy Implications

Despite the opportunities and challenges that big data analysis poses for human development, as of yet, there remains limited empirical research into these issues in contexts of development. This book is among the first of its kind in exploring the latest theoretical, practical, and ethical advances at the intersection of big data and human development. Through this unique focus and contribution, the book is outstanding in its position to examine current policy related to the adoption of data-intensive approaches to tackle development challenges and offer policy advice based on carefully collected and synthesized evidence.

Research Frontier of Big Data and Development

Even though big data in development is an emerging field, a handful of academics, institutions and projects have built critical foundations of theoretical and practical knowledge on the topic. This book is unique in gathering many

of these pioneering voices into one volume and in introducing the reader to the frontier of research and practice on the topic. Beyond addressing the theoretical considerations of development work and research taking place in the context of a datafying world and including a number of case studies and practical examples of the application of big data analysis in the context of development, the volume concludes with a section on the ethics, justice, and legal frameworks related to big data and development. This emerging research field is of paramount interest to those working with big data in development, as it illuminates how the measurement and analytical choices made in using big data analysis in development are far from being neutral or value-free considerations.

Accessible Account of Methodological Approaches

While big data has become a popular term in the vocabulary of academics, practitioners, and policymakers working with issues of development, the tools and approaches to meaningfully and effectively obtain, manage, and analyze such data are much less frequently part of their toolboxes. This book bridges this gap through presenting a selection of key methodological approaches for big data analysis in development and in particular in the context of participatory projects and initiatives. In particular, the book covers such themes as data acquisition, data management, data mining and statistical analysis, network science, visual analytics, and geographic information systems and discusses them in light of practical applications in development projects and initiatives.

How to Approach this Book

The book is divided into four main parts and further into chapters and subsections. These levels of organization have clearly defined topics and research questions and follow appropriate text structures making the content easy to understand. The book can either be read from front to back, or the reader can approach each section as a collection of related chapters or study it chapter by chapter, as they are formatted to be free standing pieces.

Each chapter is followed by a bibliography of any references cited in the chapter. In the digital version of the book, each chapter is additionally accompanied by an abstract summarizing the chapter's focus and findings, relevant keywords describing the main themes the chapter covers, and the figures can be accessed in full colour.

Goal and Mission of the Volume

While the world has enjoyed a significant reduction in poverty over the last decades, serious development challenges remain in many sectors, in particular related to the 17 goals making up the 2030 Agenda for Sustainable Development. Although the speed of technological change has delivered an unforeseen disruption around the world over the same period, its reach has been uneven and led to inequalities in the spread of benefits and adverse effects between those who effectively harness valuable data and others who overlook the significance of data or undervalue the potential risks of its wider use. The traces of these inequalities have once again been laid bare by the COVID-19 pandemic, where many nations embarked on deploying big data and advanced analytic methods to manage the crisis (Taylor et al., 2020; Mehta & Shukla, 2022; Piccarozzi & Aquilani, 2022) and where the poor have been most affected (Lucchini et al., 2023).

The range of actors involved in global development work on the frontline addressing development challenges, and stand to gain much from well-managed adoption of big data analytics. As the methods to collect and manage data constantly evolve, and the frameworks governing its safe use are developed within and between countries, there is much to learn about how to do big data analysis well and reap its many benefits while avoiding the risks it introduces. Development actors who join the processes and discussions to shape the use of big data can work towards the inclusion of more equitable practices that benefit a wide set of people within global standards. It is our hope that this volume helps those interested in the field of global development to begin to grasp big data analysis and to obtain the foundational skills and knowledge to be able to introduce data-driven approaches into their own work.

Acknowledgements

This book is the result of myriad discussions, collaborations, support, and guidance of many friends and colleagues who share an interest in data and human development. First and foremost, we express our gratitude to the book's contributing authors, who prepared their chapters with great care and revised them thoughtfully despite the constraints introduced by the global pandemic.

This book was conceptualized as a result of discussions at the Symposium on Big Data and Human Development that took place in Oxford. We are thankful for the attendees in the symposium as well as for the organizers

and attendees of related conferences that helped us to refine the scope of this volume, in particular, the Digital | Economy | Africa Conference 2019 in Johannesburg: and the Development Studies Association's Annual Conference in London in 2022 and in Reading in 2023.

We would like to thank David Sutcliffe from the Oxford Internet Institute for his careful editing of the book and Adam Swallow and Karen Bunn from the Oxford University Press for their help and guidance along the way. We are also grateful for Duncan Passey from the Oxford Internet Institute for helping us with the practical side of managing the publication of the book.

Throughout this project, we have relied on the support and reflections of colleagues, close friends, and family. Sanna is grateful to her colleagues at the Oxford Internet Institute, The Alan Turing Institute, and the Development Studies Association for the insightful discussions and shared research projects. She wishes to thank Omar for his encouragement, patience, and reflections and Marjut and Jaakko for always being there. Eduardo is deeply grateful for the many forms of support he has received from his family. Mark wishes to thank all of the brilliant researchers in the Geonet and Fairwork projects he has been privileged to learn from and collaborate with.

Finally, we are thankful for Pat Gillevet, Associate Dean of Research at the College of Science of the George Mason University, and the Oxford Internet Institute at the University of Oxford for their generous funding that enabled this volume to be published under an open license. It is our hope that this book will be accessible to communities interested in and working with development data around the world and an open license helps us to share this book with a wider audience than an academic volume generally reaches. Sanna Ojanperä's time on the project was partially funded by The Alan Turing Institute's Doctoral grant TU/C/000020 and her and Mark Graham's time was partially funded by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) [grant number 335716].

References

Alalawneh, A. A. F., & Alkhatib, S. F. (2021). The barriers to big data adoption in developing economies. *The Electronic Journal of Information Systems in Developing Countries*, *87*(1), n/a. https://doi.org/10.1002/isd2.12151

Ali, A., Qadir, J., Rasool, R. ur, Sathiaseelan, A., Zwitter, A., & Crowcroft, J. (2016). Big data for development: Applications and techniques. *Big Data Analytics*, *1*(1), 2. https://doi.org/10.1186/s41044-016-0002-4

Al-Qirim, N., Tarhini, A., & Rouibah, K. (2017). Determinants of Big Data Adoption and Success. *Proceedings of the 1st International Conference on Algorithms, Computing and Systems*, 88–92. https://doi.org/10.1145/3127942.3127961

Arora, P. (2016). Bottom of the data pyramid: Big data and the global South. *International Journal of Communication*, *10*, 1681–1699.

Benjamin, R. (2019). *Race After Technology*: *Abolitionist Tools for the New Jim Code* (1st edition). Polity.

Bertoni, E., Fontana, M., Gabrielli, L., Signorelli, S., & Vespe, M. (Eds). (2023). *Handbook of Computational Social Science for Policy*. Springer International Publishing. https://doi.org/10.1007/978-3-031-16624-2

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, *3*(2), Article 2, 111–115. https://doi.org/10.1038/s42256-021-00296-0

Cox, M. and Ellsworth, D. Application-controlled demand paging for out-of-core visualization, Proceedings. Visualization 97 (Cat. No. 97CB36155), Phoenix, AZ, USA, 1997, IEEE, pp. 235–244, https://doi.org/10.1109/VISUAL.1997.663888.

Dalton, C., & Thatcher, J. (2014). *What Does A Critical Data Studies Look Like, And Why Do We Care?* https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

DataKind. (2023). *DataKind*. https://www.datakind.org/datakind.org

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings*. *1644*(1), 104. https://doi.org/10.1063/1.4907823

Digital Impact Alliance. (2023). *Principles for Digital Development*. Principles for Digital Development. https://digitalprinciples.org/about/

D'ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Dunn, J. B., & Balaprakash, P. (2021). *Data Science Applied to Sustainability Analysis [electronic resource]*. Elsevier. https://ezproxy-prd.bodleian.ox.ac.uk/login?url=https://www.sciencedirect.com/science/book/9780128179765

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089. https://doi.org/10.1126/science.1243089

Esteban, S. A., Urquía-Grande, E., de Silva, A. M., & Pérez-Estébanez, R. (2022). Big Data, Accounting and International Development: Trends and challenges. *Cuadernos de Gestión*, *22*(1), 193–213. https://doi.org/10.5295/CDG.211513SA

The Flowminder Foundation. (2023). *Flowminder—Innovative solutions from mobile data analytics*. https://www.flowminder.org

GPSDD. (2022). *Global Partnership for Sustainable Development Data*. Global Partnership for Sustainable Development Data. https://www.data4sdgs.org/

Graham, M., & Dittus, M. (2022). *Geographies of Digital Exclusion: Data and Inequality*. Pluto Press. https://cir.nii.ac.jp/crid/1130293004051685790

Graham, M., Ojanperä, S., & Dittus, M. (2019). Internet Geographies: Data Shadows and Digital Divisions of Labour. In M. Graham & W. H. Dutton (Eds), *Society and the Internet: How Networks of Information and Communication Are Changing Our Lives* (pp. 58–79). Oxford University Press.

Haug, S. (2021, September 28). What or where is the "Global South"? A social science perspective. *Impact of Social Sciences*. https://blogs.lse.ac.uk/impactofsocialsciences/2021/09/28/what-or-where-is-the-global-south-a-social-science-perspective/

Heeks, R. (2017). *Information and Communication Technology for Development* (1st edition). Routledge.

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, *34*(1), 135–174. https://doi.org/10.1111/dpr.12142

HLG-PCCB. (2017). *Cape Town Global Action Plan for Sustainable Development Data*. High-level Group for Partnership, Coordination and Capacity-Building for Statistics. https://unstats.un.org/sdgs/hlg/cape-town-global-action-plan/

Hoogeveen, J., & Pape, U. (2019). *Data Collection in Fragile States: Innovations from Africa and Beyond*. Springer International Publishing AG. https://doi.org/10.1007/978-3-030-25120-8

Horner, R., & Hulme, D. (2019). From International to Global Development: New Geographies of 21st Century Development. *Development and Change*, *50*(2), 347–378. https://doi.org/10.1111/dech.12379

Hussein, A. A. (2020). Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). *Journal of Information Security*, *11*(4), Article 4, 304–328. https://doi.org/10.4236/jis.2020.114019

IEAG. (2014). *UN Data Revolution Group*. https://www.undatarevolution.org/

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Jemielniak, D.,(2020), *Thick Big Data: Doing Digital Social Sciences*, (Oxford: Oxford University Press)

Kankanhalli, A., Hahn, J., Tan, S., & Gao, G. (2016). Big data and analytics in healthcare: Introduction to the special section. *Information Systems Frontiers*, *18*(2), 233–235. https://doi.org/10.1007/s10796-016-9641-2

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (1st edition). SAGE Publications Ltd.

Kloß, S. T. (2017). The Global South as Subversive Practice: Challenges and Potentials of a Heuristic Concept. *The Global South*, *11*(2), 1–17. https://doi.org/10.2979/globalsouth.11.2.01

Kshetri, N. (2014). The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns. *Big Data & Society*, *1*(2), 205395171 4564227.

Kshetri, N. (2016). *Big Data's Big Potential in Developing Economies: Impact on Agriculture, Health and Environmental Security*. CABI. https://doi.org/10.1079/9781780648682.0000

Kshetri, N. (2017). *Big data and cloud computing for development: Lessons from key industries and economies in the Global South [electronic resource]*. Routledge.

López-Vargas, A., Fuentes, M., & Vivar, M. (2020). Challenges and Opportunities of the Internet of Things for Global Development to Achieve the United Nations Sustainable Development Goals. *IEEE Access*, *8*, 37202–37213. https://doi.org/10.1109/ACCESS.2020.2975472

Lucchini, L., Langle-Chimal, O., Candeago, L., Melito, L., Chunet, A., Montfort, A., Lepri, B., Lozano-Gracia, N., & Fraiberger, S. P. (2023). *Socioeconomic Disparities in Mobility Behavior during the COVID-19 Pandemic in Developing Countries* (arXiv:2305.06888). arXiv. https://doi.org/10.48550/arXiv.2305.06888

Mann, S., & Hilbert, M. (2020). AI4D: Artificial Intelligence for Development. *International Journal of Communication*, *14*, 4385–4405.

Mehta, N., & Shukla, S. (2022). Pandemic Analytics: How Countries are Leveraging Big Data Analytics and Artificial Intelligence to Fight COVID-19? *Sn Computer Science*, *3*(1), 54. https://doi.org/10.1007/s42979-021-00923-y

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://nyupress.org/9781479837243/algorithms-of-oppression

OECD. (2017). *Development Co-operation Report 2017: Data for Development*. OECD. https://doi.org/10.1787/dcr-2017-en

OECD. (2021). *Development Co-operation Report 2021 Shaping a Just Digital Transformation*. OECD.

OED Online. (2022). Information, n. In *OED Online*. Oxford University Press. https://www.oed.com/view/Entry/95568

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Piccarozzi, M., & Aquilani, B. (2022). The role of Big Data in the business challenge of Covid-19: A systematic literature review in managerial studies. *Procedia Computer Science*, *200*, 1746–1755. https://doi.org/10.1016/j.procs.2022.01.375

Sagiroglu, S., & Sinanc, D. (2013). Big data: A Review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

Sud, N., & Sánchez-Ancochea, D. (2022). Southern Discomfort: Interrogating the Category of the Global South. *Development and Change*, *53*(6), 1123–1150. https://doi.org/10.1111/dech.12742

Taaxylor, L., Sharma, G., Martin, A., & Jameson, S. (Eds). (2020). *Data Justice and COVID-19: Global Perspectives*. Meatspace Press. https://meatspacepress.com/data-justice-and-covid-19-global-perspectives/

Thurston, A. (2020). *A Matter of Trust: Building Integrity into Data, Statistics and Records to Support the Achievement of the Sustainable Development Goals*. University of London. https://doi.org/10.14296/1220.9781912250356

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C. M., Ezer, D., Haert, F. C. van der, Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., Wever, W. de, … Clopath, C. (2020). AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15871-z

Uitto, J. I. (2021). *Evaluating Environment in International Development* (2nd ed., Vol. 1). Routledge. https://doi.org/10.4324/9781003094821

UN Global Pulse. (2012). *Big Data for Development: Opportunities and Challenges*. UN Global Pulse. https://www.unglobalpulse.org/document/big-data-for-development-opportunities-and-challenges-white-paper/

UN Global Pulse. (2013). *Big Data for Development: A primer*.

UN Global Pulse. (2018). *UN Global Pulse Principles on Data Privacy & Protection*. https://www.unglobalpulse.org/policy/ungp-principles-on-data-privacy-and-protection/

UN-CEBD. (2023). *UN Big Data*. https://unstats.un.org/bigdata/

United Nations. (2015). *A World That Counts: Mobilizing the Data Revolution for Sustainable Development—Report of the Secretary-General's Independent Expert Advisory Group on the Data Revolution for Sustainable Development*. United Nations. https://digitallibrary.un.org/record/3882725

United Nations. (2018). *Principles on Personal Data Protection and Privacy*. https://unsceb.org/privacy-principles

United Nations. (2023). *Big Data for Sustainable Development*. United Nations; United Nations. https://www.un.org/en/global-issues/big-data-for-sustainable-development

United Nations Development Group. (2017). *Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda*. United Nations Development Group. undg.org/document/data-privacy-ethics-and-protection-guidance-note-on-big-data-for-achievement-of-the-2030-agenda/

Unwin, P. T. H., & Unwin, T. (2017). *Reclaiming Information and Communication Technologies for Development*. Oxford University Press.

Vance, E. A., Love, K., & Awe, O. O. (2022). *Promoting Statistical Practice and Collaboration in Developing Countries*. CRC Press. https://doi.org/10.1201/9781003261148

Wang, S. J. (2018). *Big Data for Urban Sustainability: A Human-centered Perspective [electronic resource]*. Springer.

World Bank. (2016). *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1600-0

World Bank. (2021). *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1600-0

World Bank. (2023). *World Bank Country and Lending Groups*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

PART I

THE POTENTIAL OF BIG DATA ANALYSIS FOR HUMAN DEVELOPMENT

Increasing Digital Dividends or Deepening Digital Divides

1

Opportunities and Challenges of Data-driven Development

Sanna Ojanperä, Eduardo López, and Mark Graham

1.1 The Framework Underpinning the Book

Against the background of increasingly popular uptake of big data analysis across a range of institutions and organizations, the development community's interest in incorporating big data in its work, and the emerging concerns related to the safe and ethical use of big data, this book aims to bring together theoretical and practical perspectives where these three themes connect. In the nexus of these three wider topics there are a myriad of themes of interest, but in order to limit the scope of the book, we have decided to guide its focus towards the areas that would offer insight to the most immediate needs within the development community. Drawing from our own research outlined in the introductory chapter as well as from our discussions with the authors who have contributed chapters to this volume, we designed the book's contribution to be shaped by a theoretical framework. This framework seeks to investigate central topics that on the one hand give optimism for potential positive impacts of big data to be reaped by development stakeholders and on the other highlight some of the challenges or barriers that a development actor seeking to introduce big data analysis into their operations should be aware of and prepared to navigate.

In terms of the overall framework, the book presents a range of case studies and examples that discuss the central avenues through which big data can add value to development work, focusing on three main areas:

1) Enabling operational and strategic insight and improving quality of development outcomes through:
   a. generating new insights on an existing topic by adding finer detail into a previously coarse level of analysis;

Sanna Ojanperä et al., *Opportunities and Challenges of Data-driven Development*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Sanna Ojanperä et al. (2025). DOI: 10.1093/oso/9780198879145.003.0002

   b. allowing the investigation of completely new topics identified through previously unused data sources;
   c. increasing the geographical coverage and granularity of development data;
2) Enabling the analysis of real-time or near real-time events given the more frequent recording of observations characteristic of many big data sources;
3) Enabling cost-savings through exploiting the nature of big data and the automatability of analytical tasks.

The chapters also address five central challenges related to data-driven development and representing the five main categories of barriers associated with the adoption of big data:[1]

1) Representativeness and quality of big data sources (technical barrier);
2) Differing capabilities of development stakeholders to produce and consume big data analysis (a barrier relating to people);
3) Securing political commitment (organizational barrier);
4) Financial resources required to integrate big data analysis (economic barrier);
5) Regulatory and ethical aspects of big data in development (a barrier relating to external context).

While other opportunities (such as new products and business models) and challenges (such as surveillance, exploitation, and ownership or expropriation) related to big data in the context of development can be conceptualized, we have included these themes because they were the central topics that emerged from our discussions with the book's authors. The examples and cases discussed and referred to in the book's four parts cover the themes encapsulated by all seventeen SDGs and consider various industries and sectors as well as a range of institutions, constellations of development partners, and stakeholders. The cases derive from all continents and discuss all of the most common sources of big data as well as the central methodologies used to analyze it.

The increasing speed of technological change means that many technologies and phenomena could not be covered in this volume, but there is recent work in many of the more specific communities of practice discussed in the Introduction that can offer helpful perspectives. For instance, recent work on

[1] Please see the introduction for a further discussion.

SDGs finds that AI can facilitate the achievement of the majority of the targets across the 17 goals, but may also hinder the accomplishment of a third of them (Vinuesa et al., 2020). Echoing the framework of this edited volume, we note that the need for regulatory insight and a closer alignment of AI-based technologies with sustainable development are crucial aspects that are often overlooked (Vinuesa et al., 2020).

1.2 The Promises and Challenges of Data-driven Development

1.2.1 The Promises of Data-driven Development

The brief review of the body of knowledge and practice about large-scale data and data science in development presented in the introduction demonstrates that for a field already focused on data and measurement, the appeal of deriving new insight and actionable evidence from big data analysis is clear. Embarking on new methodological approaches is challenging for any community of practice but given that global development[2] is a field where partnerships are formed between organizations hailing from contexts of drastically differing resources and capabilities, it is crucial to be able to access knowledge and resources in ways that can accommodate many levels of technicality. Learning from past experience is often an effective way to transfer knowledge and therefore, the documentation of data-driven projects across various sectors can be a valuable resource to support learning about the different sources of big data and the use of the insights generated through their analysis. Establishing big data projects is often a challenge for development organizations used to analyzing more traditional data, and examples of strategies that supported successful projects can aid the planning and operations of others wishing to include data science in their development work.

However, in order for the benefits of data-intensive development to lead to improved outcomes, there is a need to access relevant large-scale data (either through identifying appropriate existing sources or collecting bespoke big datasets) and to improve the familiarity with analytical approaches both in order to execute them solidly as well as to be able to evaluate and assess data-intensive work. Reviewing examples of past projects and initiatives can assist with learning about data sources and relevant methodologies, as well as illustrate the central benefits of big data analysis. While there are many such

[2] Please see the Introduction for a more detailed discussion about why we chose to use this term.

potential benefits, we focus here on three main areas: (1) operational and strategic insight, (2) real-time or near-real time observation, and (3) cost savings. These three benefits have proven to be recurring themes in the available academic, policy, and practice-oriented literature.

1.2.1.1 Operational and Strategic Insight

One of the key benefits of incorporating big data sources and data-driven analytical approaches is the operational and strategic insight that development actors can gain. These insights can be used to address thorny development questions and may help improve the quality of development outcomes in many ways, including the following three:

(1) Data-driven analysis can introduce **new insights on a topic that has been previously studied**, for example by complementing coarse data with finer details. As an example, representative surveys are necessarily limited in the number of questions that can be included in order to ensure a high completion rate. Where survey-generated responses can be complemented with data derived for instance through less structured sources such as public discussions carried out via radio or citizen-generated feedback, it may be possible to gather insights that would otherwise not have been collected.

(2) Where novel data sources are adopted, **completely new topics might be identified** which were not detected in previously used data sources. For instance, census data can characterize the urbanization of areas of interest through monitoring population trends and paying attention to variables such as employment, education, income, and health. Where these data are complemented with data from mobile phones indicating precise location of users, relationships between them, and voice and text produced by them, urban dynamics can be studied with much more detail on the behavioral aspects of the urban population.

(3) The adoption of big data sources and especially the application of earth observation data and geolocated transaction or communication data tends to dramatically **increase the geographical coverage of development data**. Many development challenges have a geographical dimension, and being able to utilize more detailed geographical comparisons can offer important insight to help address them.

The availability of new data sources of course does not mean that older and more conventional sources of data have diminished value. Qualitative, ethnographic, and other forms of data all have their own strengths and can all yield unique insights. Big data can both fill gaps left by the limitations of other sources, and offer powerful synergies when combined with other forms of data.

1.2.1.2 Real-time and Near Real-time Observation

A particular benefit of adopting big data and data-driven or computational analytical methods is access to measurement efforts of such high frequency that they can be seen to constitute real-time or near-real time observation. Whereas most data used in development are collected by institutions over a long period of time through fairly infrequent observations (such as censuses or widely disseminated surveys) with the aim to strike a compromise between sample coverage and temporal frequency, or over a shorter period of time in order to measure the impact of certain development interventions or initiatives (such as beneficiary surveys or experimental measurements), the possibility to record real-time or near real-time observations opens up new opportunities for development actors (Vigneri, 2022). This approach may be particularly useful for areas of development work where long analytical and response timelines may risk citizen safety or wellbeing, such as disaster relief or vaccination schedules (for instance during the COVID-19 pandemic). It can also support the monitoring of macroeconomic indicators, which often rely on complex evaluation procedures leading to the data being produced and analyzed with a considerable lag (World Bank, 2021). Real-time or near real-time data arising from digitalized transactions such as prices, flows of trade and commerce, or human mobility may allow for observation of trends and patterns in these areas as they take place, avoiding lags in measurement. Relatedly, remote or other sensor-based data, such as nighttime illumination collected through earth observation, air quality indicators, or indicators related to natural disasters, or even machine-reported data on the status of a certain technical system like water pumps can assist in understanding and overseeing the condition of critical infrastructure or other important systems and facilitate the ability to respond swiftly to potential issues. Studying the role of real-time data in impacting adaptive management practices in the development sector, Barnett and colleagues (2019) note that while the data itself are important, the role of a strategic and cultural environment that is willing and able to incorporate data-driven insights cannot be underestimated.

1.2.1.3 Cost Savings

The integration of big data and data-driven or computational methods, albeit not free of cost, may lead to cost savings in various areas of a development organization. Many traditional approaches to collecting development data incur significant expenses, as running sizeable measurement efforts such as censuses or representative household surveys is relatively costly (Jerven, 2017; Kilic et al., 2017). Savings may be achieved where similar topics can be measured through big data sources or citizen-generated data for instance,

although care must be taken in assessing the quality of such alternative data sources. Further, the insight derived from the application of data-driven methods can lead to a more efficient use of resources inside the organization, for instance enabling the direction of investment in areas that are most impactful and avoiding directing scarce budgets to areas where outcomes may be less certain. Additionally, the automatability of the analytical pipeline through the use of data-driven methods may help reduce costs via decreasing the need for manual tasks and by reducing errors or security breaches where the data management pipeline minimizes the risk of human error.

While examples about integrating big data and data science are important, equally necessary is the examination of any data-intensive tools and analytical processes against the existing data and analytical landscape and interrogating where the new larger data and methods can complement existing approaches. Global development projects tend to pose questions and address themes through established methodological approaches which have been tested through years of iteration, and much can be learned from reflecting on the large data driven methodologies against this domain knowledge (Blumenstock, 2018). While the experience of one project or initiative may not transfer directly to another context, reviewing the experience of integrating big data, interrogating the value of the insights and evidence derived from its analysis, and addressing any challenges or shortcomings identified along the way can help a development actor to design their own projects.

1.2.2 Big Data Is Not a Panacea for Development: The Challenges of Data-driven Development

While big data offers many potential benefits that development actors can aim to take advantage of in their work, there are similarly important caveats and limitations that require careful consideration in order to ensure that the use of big data does not introduce any shortcomings or harms. Some of the chief dimensions that are critical to development include interrogating the data and the data science methods in terms of their appropriateness or representativeness; in light of the differences in capabilities and resources across development actors and contexts; in view of the political commitment associated with big data; regarding the financial resources that the use of data-driven or computational approaches necessitate; in terms of the regulatory aspects related to securing the data and protecting the privacy of data subjects; and through assessing the need for new approaches towards ethical considerations pertaining to big data and data science approaches to development.

1.2.2.1 Representativeness and Quality of Big Data and Data Science in Development

Despite being initially touted as a silver bullet capable of eradicating many challenges, big data analysis should not be considered a panacea for development (Blumenstock, 2018). In order for data to enable the derivation of informative and actionable insights, they need to meet high standards of quality[3] and be governed by policies and contracts that safeguard their collection and storage.[4] The new large data sources and the analytical methods devised to extract insights from them are only helpful when they accurately describe the underlying population of beneficiaries or citizens or their activities. Many topics that are relevant to development questions such as gender, disability, or migration, have been challenging to represent in more traditional development datasets that are often collected through surveys and are still poorly reflected in some of the popular big data sources. For instance, data about refugees have been traditionally difficult to collect and the availability of big data sources has done little to alleviate the problems refugees face (Angenendt et al., 2016). Where development work aims to address questions pertaining to particular groups of individuals, utilizing big data may face limitations where the feature of group membership is not included or easily deducible (Gisselquist, 2021). Big data that derive from digitally enabled sources may encode existing digital or societal divides[5] into the recorded data. For instance, big data derived from smartphone usage will not reflect those using feature phones, basic phones, or no phones at all, and any decisions taken based on such data exclude those on the wrong side of the respective digital divide. While corporate data philanthropy makes some datasets easily accessible, the nature of much of big data and data science methods originating from corporations and particularly from large tech companies introduces its own unique challenges, as data or algorithms that are produced for a particular corporate goal may not be appropriate for addressing development questions. These challenges of representativeness and appropriateness merit careful and critical scrutiny of big data sets as well as the analytical methods used, and if done well, allow us to avoid what Hilbert (2016) terms the "paradigm of development data." The field of development has long been criticized of a weakness in collecting and using data and now that there is an abundance of new big data sources to consider, data is still at risk of not

[3] For a more detailed discussion of the central dimensions of data quality such as representativeness, completeness, accuracy, availability, consistency, and timeliness, please refer to Chapter 11 by Lopéz and Ojanperä.

[4] For a more detailed discussion of data handling, including data characteristics, please refer to Chapter 10 by Lopéz and Ojanperä.

[5] With digital divides we refer to the three conceptualizations of inequalities in access and use of data as summarized by Cinnamon (2020).

being fully utilized, thus ending up in what Custer and Sethi have called "data graveyards" (2017).

Practitioners should also take care not to fall into the trap of assuming that poor quality data are better than no data. No matter how large a dataset, if it is blind to, or misrepresents key social, economic, political, or environmental trends or characteristics, analysts and decision-makers risk making harmful choices as a result of those data. It is therefore always important to explore not just what datasets tell us, but also what, who, and where they leave out.

1.2.2.2 Differing Capabilities and Remits among Development Actors and across Contexts

The key to unlocking the value of big data in development—that is, capacity for data analytics—is not evenly spread between development actors, and varies largely between countries (Orhan & Guajardo, 2022). There are significant inequalities both between and within countries in terms of directly related skills and capacities such as data literacy, statistical capacity, infrastructure to access, collect, store, and process data, facilities for cloud computing, the presence of data-generating businesses (in particular with platform-oriented business models), in addition to the availability of relevant big data sets, as discussed in the section above (World Bank, 2016). Whereas larger development actors may have entire departments focused on research which can concentrate on producing more research-oriented outputs, many smaller development actors operate through working on more narrowly defined projects and face pressure to deliver results on shorter timescales and do not often have as much analytical capacity in-house. Therefore, larger organizations may be better positioned to incorporate big data and computational methods while smaller organizations may struggle with the adoption of data-driven methods that often require lengthy periods to acquire, clean, manipulate, and mine data as well as to derive analytical insights or develop bespoke methods. The differing abilities to access large data sets relevant for development, especially where they may be siloed, under copyright, or produced by private sector entities harboring reservations about sharing their data further differentiate the potentials between prominent and less well-connected development actors to make use of big data analysis. As the use of big data analysis can exacerbate these divides through its benefits accruing to those with capacities to access, integrate and use it, it is important to ensure that knowledge about big data analysis is included in existing efforts related to strengthening statistical capacity and data literacy.

Global development is a diverse field with activities spanning across academia, public entities, private sector partners, civil society organizations,

and international organizations. Given the fragmented nature of development work and research, the entities and individuals who work in or for development include varying profiles and among those, varying attitudes towards and abilities to handle diverse data. National statistical systems are perhaps the obvious central players in devising appropriate statistical instruments to measure relevant phenomena, carrying out data collection, and producing analysis, but these agencies tend to be underfunded and understaffed. In many countries, a significant part of the available data are produced by the private sector but using these data for development purposes is challenging due to difficulties in accessing them and because repurposing them to answer questions about development may not be straightforward. While data collected by development partners and international organizations are often easier to access, they may have been collected for purposes of internal monitoring and may not be well aligned with local development questions. In order to use big data effectively there is a need to understand differences in capabilities and resources between the constellation of development actors. The World Bank (2021) recently proposed the establishment of "integrated national data systems" to address some of these shortcomings, but setting up successful integrated systems does require a good grasp of the strengths and weaknesses of the actors making up the ecosystem.

1.2.2.3 Political Commitment to Integrate and Use Big Data

A development stakeholder seeking to adopt big data in their work will need to secure political commitment to introduce a new analytical approach. Having internal support (such as from management, colleagues, or funders) for adopting big data analytics will help navigate potential challenges related to disruptions to existing processes and ensure that the new data and analytics dovetail into existing data and analytical approaches. An environment where data are highly valued is more likely to be motivated to address such challenges in order to reap the benefits of more detailed data sources. On the contrary, where political support for big data analysis is lacking, the likelihood of introducing a new analytical approach and addressing any challenges on the way becomes significantly more difficult, leading to foregoing the potential benefits. The political attitudes towards data-driven development and its contribution to evidence-based decision-making vary between countries. OECD (2017) notes that in many low- and middle-income countries (LMICs)[6] there is a "vicious cycle" where the data collected are of low quality

[6] We use the term low- and middle-income countries (LMICs) to refer to the countries that fall into the economic classification of low-income, lower-middle-income, and upper-middle-income as defined

and the key stakeholders have little interest in improving the data systems. Indeed, the authors note that weak data systems coupled with a low valuation of big data analysis are especially likely in countries suffering from fragility, less developed economies, and in small island states. In contexts where wider interest, ownership or political support for big data or deeper data reform is lacking, but development actors are nonetheless determined to explore the adoption of these methods, partnerships between stakeholders may prove a fruitful avenue to explore.

1.2.2.4 Financial Resources Required to Integrate Big Data Analysis

Beyond securing political support, those seeking to adopt big data in their development work will need to secure sufficient financial resources to support a data-driven approach. While hiring a data scientist and gaining access to a relevant data set may constitute the initial steps, adopting big data analysis as a central part of one's operations requires provisions for designing an appropriate data infrastructure, securing sufficient computing facilities, and retraining staff on the value of data. It is challenging to measure the costs related to big data analysis, as the required infrastructure, human resources, and any costs associated with obtaining the data vary greatly between projects and programs, but revising the costs related to the SDGs can provide some reference. The 2030 Agenda is often referred to as a data-driven program for action since the SDGs are all about tracking and evaluating development outcomes through indicators and targets conceptualized as quantitative measures. However, being able to evaluate progress requires having access to data points (a fact that the previous global development agenda aiming to reach Millennium Development Goals (MDGs) sorely underestimated).[7] Indeed, when the SDGs were adopted, the international community recognized the lack of statistical instruments to measure progress over many of the elements of the SDG measurement framework within many of the countries (Jerven, 2017). Establishing new measurement tools and starting novel data collection requires financial resources, and the "data funding gap" for producing data for SDG indicators on tier 1 and 2 of the framework has been estimated at USD 2.8–3.0 billion per year up to 2030 (OECD, 2017). Perhaps the shortcomings of the MDGs have inspired a more earnest focus on tracking the progress in measurement as well as in

by the World Bank (2023). Please see the Introduction for a more detailed discussion about why we chose to use this term.

[7] The Millennium Development Goals (MDGs) were adopted at the United Nations Millennium Summit in 2000 and comprised eight development goals, which were to be fulfilled by 2015. The MDGs were conceived of as an ambitious monitoring exercise covering 21 targets. They were however plagued with a number of measurement and conceptual challenges and the set targets were not met in many countries. See for example Waage et al. (2010) for a broad overview.

the achievement of the 17 areas that the SDGs address: the United Nations publishes an annual SDG Report that monitors the level of achievement[8] while the World Bank has launched a set of new statistical performance indicators to better monitor countries' statistical performance (Dang et al., 2021). These efforts and the wider discourse taking place around SDGs highlight that data are perceived to play a key role in meeting development commitments. They also illustrate that despite the increased presence of data, their collection and use are rarely free. While some types of big data may alleviate some of the cost issues related to traditional development data such as household surveys, it is important to carefully plan what financial resources are needed for big data projects in order to ensure they can be carried out sustainably.

1.2.2.5 Regulatory and Ethical Considerations Pertaining to Big Data and Computational Approaches to Development

Many of the topics and themes central to global development touch on sensitivities and vulnerabilities of parts of the population that are already at a disadvantage. Recording and using data that feature personal and sensitive information necessitates consent and requires particular care to manage safely and securely. The nature of many big datasets containing very detailed information about individual preferences, expressed language, social interactions, behaviors, activities, financial transactions, and mobility patterns (to name a few) opens up further needs for ensuring privacy and security of the data subjects. Where data from multiple sources are combined, the risks of deanonymization must be handled carefully in order to avoid harm for the data subjects (Backstrom et al., 2007). While LMICs may be at varying stages of devising and implementing the regulatory and policy frameworks governing the processes to handle data, there have been calls among development actors to ensure that data processing methods, organizational processes to manage data, and interoperability between organizations handling data are updated and reflect high standards of security (OECD, 2021). Some development stakeholders have proposed general guidelines to manage big data, such as the United Nations Development Group's (2017) guidance note for big data that focuses on data privacy, ethics, and protection. Chapter 10 of this volume, on managing data, gives a brief overview of some of the central considerations that a practitioner managing a data-driven project will benefit from.

While the legal, regulatory, and policy frameworks governing the management of data are the linchpin for the safe and secure management of big data

[8] For the latest available report, see United Nations (2022).

and data science in development, they are not in place across all contexts where development work happens and will take some time to develop, adapt, and harmonize. It is important that even where these governance frameworks are not fully formed, development actors working with big data and computational approaches pay particular attention to securing the consent of data subjects and take measures to protect their privacy. Communities of practice that cross-cut academia and development work have emerged as sites fostering discussions about the critical dimensions of data use and management. While a critical discussion about data should be encouraged within any data ecosystem, the insights from ethical guidelines and frameworks developed by communities of practice are particularly helpful for contexts where the formal institutions governing data are nascent. The concept of data justice focuses on the power relations underpinning today's increasingly datafied societies and examines the social consequences of big data and related technologies (Directorate-General for Parliamentary Research Services (European Parliament) et al., 2022; Heeks & Renken, 2018; Dencik, 2022). These resources and the underpinning debate can greatly inform anyone planning to work with big data in global development and who is interested in thinking through the implications. In order to translate the insights from critical data studies and other related communities of practice to global development, examples discussing the particular context of development practice are in high demand.

1.3 Insights from this Edited Volume

While it is impossible to summarize the insights from all of the book's chapters within this introductory chapter, we think that it is useful to briefly comment on how the chapters relate to the framework of the book and therefore how they contribute to addressing the wider analytical focus the book adopts. While most chapters touch on all of the themes of the framework, below we raise points from a few of the chapters that discuss each theme in greater depth or approach it from complementary viewpoints.

1.3.1 Illustrating the Benefits of Big Data

The structure of the book is designed to respond to one of the central needs identified for successful integration of big data and data science in development: the transfer of knowledge through examples and case studies that draw from various contexts relevant to development. Through the four parts of the

book, the reader is introduced to applications of big data and development in all continents, touching on the themes of all of the 17 SDGs, covering a range of sectors and topics, spanning across the key sources of big data, and utilizing all of the central methodological approaches suitable for big data analysis. The examples involve a diverse set of development stakeholders and researchers working in the public sector, academia, private entities, and civil society organizations.

The three areas identified as some of the central benefits derived from big data analysis, namely, (1) increased operational and strategic insight and higher quality of outcomes owing to additional granularity, allowing both the derivation of new detail and the emergence of new themes, and improving geographical coverage; (2) increased temporal frequency; and (3) cost savings, are approached from a multiplicity of viewpoints throughout the book.

1.3.1.1 Operational Insights

Among the examples illustrating how big data can lead to benefits related to operational, strategic, and quality aspects of development work, in Chapter 7 Lokanathan, Perera-Gomez, and Zuhyle describe how big data analysis can add new detail into an already studied phenomenon. The authors look at how including big mobility data in the prediction of the spread of dengue fever in Sri Lanka helped to produce more precise estimates. The study derived variables depicting mobility of individuals from a large mobile network dataset and included them into a model of dengue fever incidences, which resulted in improved predictions. In Chapter 8 Hidalgo-Sanchis notes that radio is the main media platform in Uganda providing citizens with the principal channel for voicing feedback and opinions. She details a project where the transmission from community radio stations is recorded and processed into text-based data in Luganda and Acholi, two vernacular languages of Uganda. Content analysis was applied on the dataset, which allowed the extraction of critical insights such as citizen testimonials and public opinions about topics such as quality of public service delivery and public sector accountability that could be used to support the design and monitoring of policies. Highlighting a related benefit of big data—allowing the emergence of new themes and insights—Heeks notes in Chapter 13 how data initiatives by virtue of creating "data doubles" have altered the "imaginaries" of those involved, as well as the wider discourse about the respective phenomena. He describes how introducing a data-based system to a bus transportation company in India made visible the operations the bus fleet and the crews, which in turn allowed the managers of the company to gain a much more detailed view into the firm's operations. Looking at the insights that big data can generate

from a more theoretical point of view, Letouzé outlines three functions for data in Chapter 4. He notes that big data can be used as a lens (to more accurately describe and quantify phenomena), a language (where data literacy and statistical literacy enable individuals to participate in meaningful discussions about data and data-driven methods), and a lever (where data are used as a transformative tool to enable individuals and communities to arrive at and implement decisions with measurable targets), and employing each of the three functions is likely to yield optimal outcomes. In discussing another benefit of big data, adding geographic detail, in Chapter 3, Ndemo notes the MapKibera project, which used spatial tools to create an interactive digital map of Kibera, a populous shanty in Nairobi. The exercise allowed the identification of previously unknown details about the shanty, which enabled the planning of more targeted services for the underserved population. Chapter 11 by López and Ojanperä illustrates how diverse sources of big data have been used to generate novel research findings spanning all 17 SDGs.

1.3.1.2 Temporal Detail

Increasing temporal detail is among the most important benefits of big data and Oliver illustrates this in two cases that she discusses in Chapter 2: early warning for floods in Mexico and detecting hotspots for crime in the United Kingdom. In Mexico she describes her project that analyzed the social reaction to floods through mobile data and which enables a quick, near real-time identification of areas that are most affected by floods. In the UK, her research involved data about human activity and demographics in order to detect the emergence of crime hotspots in the city, supporting the authorities in planning how to best respond to crime. In Chapter 5, Njuguna and McSharry construct a multi-dimensional poverty index for Rwanda utilizing large datasets as proxies for disposable income, such as mobile telephone ownership and usage derived from call detail records and the nighttime illumination recorded through satellite data. They are able to predict poverty on the regional level in Rwanda and note that estimates such as theirs may assist with providing information about poverty levels which are otherwise measured in the Rwandan census in ten-year intervals.

1.3.1.3 Cost Savings

Cost savings derived from the use of big data analysis is an attractive prospect in the context of development work, where domestic resources are limited and highly competed for, and the budgets of development cooperation are constrained. In Chapter 5, Njuguna and McSharry note that using big data sources to construct a multi-dimensional poverty index allows increasing

the frequency of poverty measurements without having to assign a significant budget to collecting comparable data through household surveys. In Chapter 6, Lepri, Oliver, and Pentland discuss the Open Algorithms program (OPAL), which involves a platform that holds private sector data securely and allows certified users to access it for the purposes of initiatives fostering social good. They note that such a platform would enable data owners to monetize their data by establishing more direct links between data owners and monetary fees to access data through the platform and through enabling charging micro-payments that charge small transaction costs.

In the examples given in this book, big data sources are most often used complementarily with existing data in order to validate the data-driven approaches, but also to better integrate new data into areas where a substantial body of more traditional quantitative work has already been carried out.

1.3.1.4 Practical Resources for the Aspiring Data Scientist in Development

Beyond these brief references to the topics discussed by the contributing authors, the three chapters in Part III of the book present the reader with a more practical viewpoint into data-driven development. Chapter 10 by López and Ojanperä discusses the considerations and processes underpinning safe and effective research with big development data and offers a tutorial on data management. Chapter 11 by the same authors reviews a diverse range of examples of big data analysis of a variety of sources integrated into development and offers examples connecting those data sources to all of the 17 SDGs. In Chapter 12, López and Ojanperä offer an overview of the analytical approaches to extract meaning from big data in development and relate this discussion to the skills and training landscape of development professionals. The three chapters in Part III of the book are accompanied with plentiful references to source materials that the aspiring data scientist interested in development research questions will likely find useful.

1.3.2 Characterizing the Challenges and Barriers of Adopting Big Data

The five barriers and challenges related to adopting big data discussed in this book relate to: (1) the representativeness and quality of big data and data science in development; (2) the differing capabilities and resources across development actors and contexts; (3) political commitment associated with big data; (4) costs associated with big data; and (5) regulatory and

ethical considerations pertaining to big data and data science approaches to development. The chapters in this volume touch on these five areas from a multiplicity of viewpoints.

1.3.2.1 Representativeness and Quality of Big Data and Data Science in Development

The success of data-driven development work hinges on the quality and appropriateness of big data sources for their intended application in development. In Chapter 5, Njuguna and McSharry discuss the challenges of working with big data deriving from private sources when they describe their proof-of-concept project that constructs a poverty index for Rwanda from big mobile and satellite data. Constructing the index with big data sources that proxy disposable income allows near real time monitoring of poverty, which otherwise would be measured at large intervals. The authors use data from the leading mobile network operator (MNO), which has a broad coverage in the country, but focuses on urban areas. While the authors describe a proof-of-concept tool that is able to retrieve the poverty estimates achieved by a more traditional data collection method, they note that ideally the tool would encompass data from all of the MNOs so that mobile phone activity in the country would be represented as accurately as possible. In Chapter 6, Lepri and colleagues focus on algorithmic decision-making and note that such processes may well include bias and discrimination and may suffer from a lack of transparency and accountability if not designed carefully. They mention that in order to avoid such a "tyranny of data" or "tyranny of algorithms" and instead support transparency and fairness, a multi-stakeholder approach involving researchers, practitioners, and policymakers should be deployed to create and analyze algorithmic decision-making processes. The integration of big data and related analytical methods to development introduces challenges related to quality and appropriateness—however, where they are addressed well they will, big data can offer valuable contributions to addressing development questions.

1.3.2.2 Differing Capabilities and Resources among Development Actors and across Contexts

The uneven spread of capacities and resources to take advantage of big data in development is one of the leading challenges, limiting its efficient adoption. In Chapter 7 Lokanathan, Pereira-Gomez, and Zuhyle describe the journey of setting up the big data think tank, LIRNEasia, in Sri Lanka, where some of the key limitations have to do with a shortage of data scientists and researchers with strong computer science skills. They found themselves

competing over the small pool of locally based skilled professionals with private firms offering attractive salary packages. The authors describe that some of the ways in which they addressed the situation included forming multi-disciplinary teams which alleviated the need for computing skills, introducing robust organizational record-keeping to ameliorate the inevitable churn of personnel, and continuously upgrading the research capacity in-house as well as in the education and training institutions that could in turn provide the think tank with collaborators and colleagues. LIRNEasia was instrumental in founding a multi-disciplinary research center, DataSEARCH, in a leading university in Sri Lanka as well as the Citra lab, the country's first social innovation lab. Writing about their experience working in the Council for Scientific and Industrial Research (CSIR), the primary research arm of the South African government, Marivate and Moorosi note in Chapter 9 that among the main obstacles they faced when designing data science projects and tools for various areas of the government was handing the tool or project over to an entity that might not have the right skills to use and maintain it. They responded to the challenge by expanding the education and training offer through various avenues. The CSIR founded the Data Science for Insight and Decision Enablement (DSIDE) program, which would select a cohort of students to work at the CSIR on data study projects proposed by various government departments for three months each year. The selection process was explicitly designed to produce a diverse cohort with students from different universities in South Africa, and the training provided observed varying levels of skills. The CSIR additionally founded a summer school focusing on data science, AI, and machine learning, and supported various universities with designing programs and courses. They were also involved in organizing events aiming to build a wider community around the topics in the country. While not every project or initiative seeking to apply big data analysis can put in place such comprehensive measures to develop their own resources and the capacities of those around them, many of the other chapters in the volume describe how differences in capacities were successfully addressed by forming partnerships between organizations and entities that enabled setting up a successful project. In Chapter 4, Letouzé theorizes about the trajectory of achieving progress with respect to certain development targets and characterizes the process taking place with the term "learning loops." In this view, progress towards a development target is likely not linear over time, but goes through more and less productive periods, which in themselves help to inform the process. He notes that being able to derive value from data through the three functions he outlines—using data as a lens, as a language, and as a lever (see section 1.3.1.1. for a description of the terms)—enables deriving analytical insights and acting upon them therefore

supporting the learning and decision-making processes along the non-linear path illustrated by the "learning loop."

1.3.2.3 Political Commitment Associated with Big Data

Implementing successful data science projects in development requires addressing contextual factors such as securing political commitment, finding relevant partners, and developing the analytical offering in a situation where there are scarce resources, and the value of new methods may not be appreciated or understood. In Chapter 3, Ndemo investigates policymakers' attitudes to big data in Africa and notes that investment into big data remains low in the region with decision-makers harboring skepticism and hesitancy about open data in particular. He shares his personal experience in drafting an "Access to Information" bill and seeing it through to being adopted as an Act of Parliament during his time as a former Permanent Secretary of Kenya's Ministry of Information and Communication. The process took 10 years with delays owing to fears about exposure to questionable activities. He observes that many African policymakers agree on the potential benefits of big data and posits that "to carry society forward, our data analytics must be big, open, and transparent." He describes the region as continuing to invest in data centers and other data infrastructure but expects that adopting data-driven decision-making will take time and the measurement exercise to reach the SDGs may be carried on by future generations of leaders. In Chapter 9, Marivate and Moorosi highlight the difficulty of not only attaining political commitment but retaining it during a lifetime of a big data project. They note that the personnel churn in the policy roles in South Africa is significant and often necessitates rebuilding relationships to achieve a deliverable. Challenges may arise particularly where a project champion who joins mid-way comes with their own agenda points, which may significantly complicate the project. The authors therefore highlight the importance of securing political buy-in from stable champions and note that political support particularly facilitates data access and helps to navigate through technical and institutional challenges. Being able to communicate the value of data-driven approaches to political stakeholders and securing stable political commitment remain key challenges for adopting big data analysis in development.

1.3.2.4 Costs Associated with Big Data

While utilizing big data may lead to savings in certain areas, there are also costs associated with adopting data analytics as part of a development actor's work, especially in first setting up the infrastructure required for effective and secure data analysis and in cases where bespoke data collection methods need to be devised. In Chapter 10, López and Ojanperä discuss a framework for effective data management and note that storing and using data safely

and securely throughout the data life cycle requires access to hardware and software that often are not free of cost. In Chapter 9, Marivate and Moorosi mention that a challenge they faced in developing a data science project for a government department was getting budgets and resources approved within the planned schedule. The authors found a solution in involving an educational institution in the project to cover some of the costs through exchange for training opportunities offered by the project, but caution that this model introduced a pricing problem further ahead when the government department was not able to match the funding to allow for the educational partner to exit the program as planned. In Chapter 7, Lokanathan, Pereira-Gomez, and Zuhyle discuss managing the costs of big data in development projects and note that LIRNEasia has been able to secure funding for their activities through networking with like-minded communities and through the advisory council they have established to oversee their research activities. Many authors in this volume similarly note the value of partnerships that can support their activities computationally. Finding ways to sustainably cover the costs of big data projects in development are prerequisites for adopting big data analysis. The chapters in this volume offer many perspectives and strategies for navigating these aspects and developing successful projects and initiatives.

1.3.2.5 Regulatory and Ethical Considerations Pertaining to Big Data and Computational Approaches to Development

Adopting data-driven approaches to development work is made more challenging by the current state of the legal and regulatory landscape governing the use of data as well as the ethical frameworks developed to guide the fair use of data. Many LMICs are in the process of developing their data policies and cybersecurity and privacy laws, resulting in a legal, regulatory, and policy landscape that is fairly varied between countries. Regionally harmonized approaches are being developed with a recent example being the African Union Data Policy Framework (2022) that aims to establish a common vision, principles, strategic priorities, and key recommendations to guide the region's countries in developing their national data systems and capabilities to effectively use and derive value from data. Given this changing context, those seeking to introduce data-driven projects in development may benefit from drawing from parallel work carried out in other communities, such as the work related to critical data studies. Relevant communities of practice comprise a diverse set of viewpoints and are active in developing ethical frameworks and guidelines that help those interested in implementing data-driven projects achieve them safely. Conceptual, theoretical, and action-oriented work that has shaped concepts and research areas such as data governance, data justice, data colonialism, and data capitalism has recently

introduced actionable insights that can directly benefit those seeking to initiate data-driven development projects. In Chapter 13, Heeks discusses and revises one of the proposed structural data justice models drawing from his experiences of field research in India and Kenya. The revised model that Heeks proposes incorporates several kinds of data justice in order to derive a "Data-Justice-for-Development Manifesto," which he notes can be used to guide development policy and practice. In Chapter 14, Taylor examines the notion of data as public good from the perspective of private sector data holders. She notes that the "responsible data" movement that develops guidelines and frameworks to derive ethical principles for data sharing has not gained much traction within corporate data holders, in particular mobile providers. She assesses the power dynamics and differing incentives that actors face and examines the two models of data sharing: first where an intermediary holds and controls access to various datasets and second where private sector entities manage the data sharing process. She notes that recent advances in implementing either model have contributed to creating a new field of data ethics, deepening the understanding of the role that privacy and ethics play in data-driven work in development. Lepri, Oliver, and Pentland discuss a model for data-sharing that falls into the first category described by Taylor in Chapter 6. They suggest that this approach offers a new dimension to the social contract, as it involves co-designing the ways to query included datasets such that they serve the needs and meet the standards of local users, for instance through avoiding biases and unintended discrimination while ensuring privacy. Several of the chapters in this volume illustrate the need for, and challenges concerning, protecting the privacy of the data subjects and the chapters by Taylor and Heeks underline the importance of obtaining their consent for data collection. Most of the chapters touch on the ethical and regulatory issues in the context of the particular projects or phenomena they discuss and present perspectives into a central dimension of data-driven development, where the legal, regulatory, and policy frameworks remain in active state of development.

References

African Union. (2022). *AU Data Policy Framework*. African Union.

Angenendt, S., Kipp, D., & Koch, A. (2016). *Many refugees, poor data: Development cooperation requires higher-quality data* (No. 7; SWP Comment). Stiftung Wissenschaft und Politik -SWP- Deutsches Institut für Internationale Politik und Sicherheit. https://www.ssoar.info/ssoar/handle/document/48078#

Backstrom, L., Dwork, C., & Kleinberg, J. (2007). Wherefore art thou r3579x? Anonymized social networks, hidden patterns, and structural steganography. *Proceedings of the 16th International Conference on World Wide Web*, 181–190. https://doi.org/10.1145/1242572.1242598

Barnett, I., Hernandez, K., Ramalingam, B., Levy, A., Oppenheimer, C., & Valters, C. (2019). Can ICT-enabled real-time data contribute to adaptive management in development programming? *Development in Practice*, *29*(3), 287–299. https://doi.org/10.1080/09614524.2018.1557596

Blumenstock, J. (2018). Don't forget people in the use of big data for development. *Nature (London)*, *561*(7722), 170–172. https://doi.org/10.1038/d41586-018-06215-5

Cinnamon, J. (2020). Data inequalities and why they matter for development. *Information* Technology for Development, *26*(2), 214–233. https://doi.org/10.1080/02681102.2019.1650244

Custer, S., & Sethi, T. (2017). *Avoiding Data Graveyards: Insights from Data Producers & Users in Three Countries*. https://www.aiddata.org/publications/avoiding-data-graveyards-insights-from-data-producers-users-in-three-countries

Dang, H.-A. H., Dinc, M., Dias, J. Z. J., Maeda, H., Pullinger, J. J., Serajuddin, U., Stacy, B. W., & Wolde, D. K. (2021). *Measuring the Statistical Performance of Countries: An Overview of Updates to the World Bank Statistical Capacity Index* (No. 157540; Technical Note). World Bank Group. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/815721616086786412/Measuring-the-Statistical-Performance-of-Countries-An-Overview-of-Updates-to-the-World-Bank-Statistical-Capacity-Index

Dencik, L. (2022). *Data justice* (1st ed.). SAGE.

Directorate-General for Parliamentary Research Services (European Parliament), Lopez Solano, J., Martin, A., Souza, S. de, & Taylor, L. (2022). *Governing data and artificial intelligence for all: Models for sustainable and just data governance*. Publications Office of the European Union. https://data.europa.eu/doi/10.2861/915401

Gisselquist, R. (2021, February 17). *A "data revolution" for sustainable development leaves gaps on inequality*. UNU-WIDER. https://www.wider.unu.edu/publication/%E2%80%98data-revolution%E2%80%99-sustainable-development-leaves-gaps-inequality

Heeks, R., & Renken, J. (2018). Data justice for development: What would it mean? *Information Development*, *34*(1), 90–102.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development* Policy Review, *34*(1), 135–174. https://doi.org/10.1111/dpr.12142

Jerven, M. (2017). How much will a data revolution in development cost? *Forum for Development Studies*, *44*(1), 31–50. https://doi.org/10.1080/08039410.2016.1260050

Kilic, T., Serajuddin, U., Uematsu, H., Yoshida, N., & Uematsu, H. (2017). *Costing Household Surveys for Monitoring Progress Toward Ending Extreme Poverty and Boosting Shared Prosperity* (SSRN Scholarly Paper No. 2905440). https://papers.ssrn.com/abstract=2905440

OECD. (2017). *Development Co-operation Report 2017: Data for Development*. OECD. https://doi.org/10.1787/dcr-2017-en

OECD. (2021). *Development Co-operation Report 2021 Shaping a Just Digital Transformation*. OECD.

Orhan, C. C., & Guajardo, M. (2022). Analytics in developing countries: Methods, applications, and the impact on the UN Sustainable Development Goals. *International Transactions in Operational Research*, *29*(4), 2041–2081. https://doi.org/10.1111/itor.13018

United Nations. (2022). *The Sustainable Development Goals Report 2022*. https://unstats.un.org/sdgs/report/2022/

United Nations Development Group. (2017). *Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda*. United Nations Development Group. undg.org/document/data-privacy-ethics-and-protection-guidance-note-on-big-data-for-achievement-of-the-2030-agenda/

Vigneri, M. (2022, March 1). *Timely evaluation in international development* [Monograph]. Centre for Excellence and Development Impact and Learning. https://doi.org/10.51744/cmwp7

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-019-14108-y

Waage, J., Banerji, R., Campbell, O., Chirwa, E., Collender, G., Dieltiens, V., Dorward, A., Godfrey-Faussett, P., Hanvoravongchai, P., Kingdon, G., Little, A., Mills, A., Mulholland, K., Mwinga, A., North, A., Patcharanarumol, W., Poulton, C., Tangcharoensathien, V., & Unterhalter, E. (2010). The Millennium Development Goals: A cross-sectoral analysis and principles for goal setting after 2015. *Lancet (London, England)*, *376*(9745), 991–1023. https://doi.org/10.1016/S0140-6736(10)61196-8

World Bank. (2016). *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1600-0

World Bank. (2021). *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1600-0

World Bank. (2023). *World Bank Country and Lending Groups*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

2

Big Mobile Data for Social Good

Opportunities and Challenges

Nuria Oliver

2.1 Introduction

We live in a world of data, of Big Data, whose volume is increasing exponentially, to the point that it is difficult to comprehend how much data we are talking about. In 2022 alone, it is estimated that the volume of data created, captured, copied, and consumed in 2022 reached 97 zettabytes of data, according to Statista, that is, 10 to the power of 21 bytes or one trillion gigabytes of data (Taylor, 2023).

Where does all this data come from? This explosion of data is the result, on the one hand, of the digitization of the physical world, particularly driven by the ubiquity of mobile phones and the exponential growth of all types of sensors—many of which are part of what is known as The Internet of Things or IoT—such that our actions in the physical world generate a digital trace or footprint (e.g., traffic sensors, credit card transactions, etc.). On the other hand, this data is also the result of our increasingly intense activity in the digital world, fueled by the massive adoption of smartphones which are connected to the Internet, with thousands of mobile apps that also leave a digital trace as a result of their use.

Although there are large volumes of data that are generated by activity without a direct link to human behavior (e.g., data collected by particle accelerators, space telescopes, environmental measurements, etc.), a significant part of this data is generated as a result of human activity, the activity of each of us.

Hence, for the first time in the history of humanity since we have existed as a species, we can measure human behavior on a large scale, which has led to the birth of a new discipline called computational social sciences (Lazer et al., 2009). Where we go, what we consume, who our friends are, how much

Nuria Oliver, *Big Mobile Data for Social Good*. In: *Human Development and the Data Revolution*.
Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Nuria Oliver (2025).
DOI: 10.1093/oso/9780198879145.003.0003

money we spend and on what are some aspects of our behavior that we can capture digitally and hence are susceptible to be modeled computationally. These data-driven models allow us to personalize our interactions with technology or with specific services. Thanks to personalization algorithms trained with our data (namely what we have consumed and liked in the past) we receive recommendations (Ricci et al., 2010) of relevant movies on Netflix, of potentially interesting books on Amazon, coupons with discounts for our favorite products, and ads and banners on Facebook and Google that are targeted to us, to our taste and interests.

Without a doubt, the economic impact of Big Data is paramount. According to a recent Accenture report called "Big Success with Big Data" (2017), 89% of respondents implemented a project that uses Big Data in business and believe that Big Data will disrupt business operations; and 85% of respondents believe that Big Data will drastically change how business is done. A study on "The European Data Market" by the European Commission (2021) estimated the value of the European Data Economy at 580,000 million euros in 2021, which is approximately 4% of the European GDP and represents a year-on-year growth of over 5%.

Medicine, banking, journalism, advertising, trade, transport, industrial production, agriculture, education, sports, science, to name a few, are disciplines that are already leveraging Big Data. The global big data and business analytics (BDA) market was valued at USD 240,600 million in 2021 and it is projected to surpass USD 655,000 million by 2029, with a compound average growth rate of 13.4% over the 2022–2029 period, according to a report published by Fortune Business Insights (2022).

However, there is a Big Data opportunity that is probably much less known despite its immense potential: the application of Big Data for Social Good (DataIQ, 2017), which is the focus of this chapter.

The existence of large-scale human behavioral data that is anonymized (i.e., preserves privacy) and aggregated (i.e., it is analyzed in a large scale, not individually) enables us, for the first time in our history, to help governments and relevant institutions in their decision making processes, to model and measure migrations, to quantify the impact of natural disasters, to anticipate the spread of pandemics or to automatically infer the socioeconomic status of a region.

The 17 Sustainable Development Goals established by United Nations at the end of 2015 include the concept of Big Data both to enable us to measure our progress in the achievement of such goals and to help us achieve them. In addition, the recent Global Partnership for Sustainable Development Data (United Nations, 2017) and the GSMA's Big Data for Social Good initiative

(2017) are two examples of large scale recent and global initiatives aimed at leveraging Big Data for positive social impact.

An article published by United Nations Global Pulse (2012) described the challenges and opportunities offered by Big Data when it is used for Social Good, and it proposed a taxonomy with three levels of use: to create "real time awareness," to assist "early warning systems" and to provide "real-time feedback." A later article about Big Data to prevent conflicts proposed three different functions of the use of Big Data (Mancini & O'Reilly, 2013): descriptive (e.g., maps), predictive (to generate predictions about the future) and prescriptive (to be able to carry causal inferences). Most of today's examples of Big Data for Social Good belong to the first two categories. However, examples of using Big Data with a prescriptive function will probably be more prevalent in the future.

My research experience in this topic started during my time as Scientific Director at Telefónica I+D (2007–2016), where I created and led a research area devoted to Big Data for Social Good, in collaboration with researchers from my team and with external institutions such as the United Nations, the Mexican Government, the World Food Program, the Bruno Kessler Foundation (FBK), MIT and Data-Pop Alliance. Next, I established this research area at Vodafone during my time as Director of Research in Data Science (2017–2019). Today, I continue to be active in this research area from my roles as Chief Scientific Advisor at the Vodafone Institute, as Chief Data Scientist at Data-Pop Alliance and as Director of the ELLIS unit devoted to Human-centric Artificial Intelligence in Alicante, Spain.

Throughout these years we empirically validated the value of data captured by the mobile network infrastructure to automatically identify the most affected areas by seasonal flooding (Pastor-Escuredo et al., 2014), to quantify the impact of government interventions to contain the spread of an infectious disease and prevent a pandemic (Frías-Martínez et al., 2012, (Miguel Angel Lozano, 2021)), to infer the socioeconomic level of a region (Soto et al., 2011), and to detect crime hotspots in London's metropolitan area (Bogomolov et al., 2014), a project which won the Datathon for Social Good (2013) that Telefónica organized in collaboration with MIT, the London Campus Party and the Open Data Institute in 2013. In this chapter I will describe two of these projects, followed by my insights and a discussion on existing barriers that we need to overcome before we can realize the tremendous potential of this opportunity.

Before describing these two projects, note that in the past 10 years there have been several data-sharing competitions where aggregate and pseudoanonymized mobile data have been shared with the research community

for Social Good purposes. These data-sharing experiences have resulted in hundreds of projects and publications. I would highlight the two Orange D4D challenges in 2013 and 2015 where they shared aggregated data from Ivory Coast and Senegal (2015) with tens of research teams worldwide, resulting in tens of projects where Big Data has been used to understand urban transport, improve official statistics, contribute to public health, and identify relevant patterns between human behavior and natural disasters, energy needs and agriculture, among others.

Telecom Italia also organized in 2014 a Big Data analysis challenge (Telecom Italia, 2014) which attracted over 1000 participants from 20 different countries which proposed 100 innovative ideas using Big Data for a variety of purposes, including social good. Other telecommunication companies such as Telenor and Vodafone are also carrying out projects on Big Data for positive social impact. Recently, Turkcell organized the Data for Refugees challenge (Salah et al., 2018), where aggregated and pseudonymized mobile data was shared with research groups to carry out research projects to shed light on the safety and security, health, unemployment, education and integration of Syrian refugees in Turkey.

From an institutional perspective, in addition to several units within the United Nations (and in particular United Nations Global Pulse and UNICEF, with whom I have collaborated), I would highlight the work of Flowminder (2023), a non-profit organization which has been working on improving the public health and the wellbeing of countries with low to medium socioeconomic status through the analysis of aggregated and (pseudo)anonymized mobile data, satellite imagery and census data; and the work of Data-Pop Alliance (2015), a non-profit alliance between the MIT Media Lab, the Harvard Humanitarian Institute, the Overseas Development Institute (ODI) and Flowminder, whose objective is to promote a people centric Big Data revolution, and where I serve as Chief Data Scientist.

In this chapter, I will illustrate the value of Big Data—particularly the data captured by the mobile network infrastructure—for Social Good by describing two projects in two different areas: natural disasters and crime. The two chosen areas have a profound impact on development both in developed and developing economies and highlight the importance of having access to real time data that would enable a timely response.

Natural disasters—such as hurricanes, earthquakes, and floods—can cause extensive damage to infrastructure and lead to a significant loss of lives, homes and livelihoods. This damage can take years to repair and can set back development progress tremendously. A study by the IMF (Bayoumi et al., 2021) found that a large natural disaster can reduce GDP growth rates by

0.7% in the year of the disaster on average and such impact can be as high as 3% in countries with low levels of disaster preparedness. Employment rates can also decrease in the aftermath of natural disasters, as workers may be displaced or unable to find work due to the destruction of businesses and infrastructure. Crime also has a significant negative impact on development by reducing the level of trust in the government and institutions, discouraging investments, and creating a negative image of the region or country. Only in the United States, the estimated annual cost of crime is USD 4.71–5.76 trillion including the direct costs of law enforcement, criminal justice and victims' losses, and the indirect costs of private deterrence, fear and agony, and time spent for avoidance and recovery.

2.2 Big Data and Natural Disasters

The worldwide economic impact of natural disasters is enormous. Only in 2022, the world has experienced 421 notable natural disasters with a financial cost of USD 313,000 million USD (4% above the 21st century average), USD 132,000 million in insurance costs (57% above the twenty-first-century average) and at least 31,300 deaths, according to the Aon report on "Weather, Climate and Catastrophe Insight" (2023). Developing countries are more vulnerable to natural disasters because of poor construction quality, weak or non-existent infrastructure for early warning and emergency relief, lack of disposable assets for emergencies and weak social safety networks to help when facing an unexpected tragedy. In addition, developing countries are amongst the most affected by climate change, which in turn leads to a higher intensity and frequency of natural disasters. According to Germanwatch's 15th edition (2020) of its global climate risk index, poor countries are the most vulnerable to the impact of climate change: 80% of migrations due to extreme climatic events in the last decade are based in Asia, where over a third of the world's poorest people live. In fact, 14 of the 15 places in the world that were the most severely affected by extreme weather events between 1999 and 2018 were developing countries.

When a natural disaster takes place, emergency relief agencies need to quickly and accurately identify the affected areas and estimate the number of individuals who have been affected by the event. In this context, Big Data and especially aggregated and pseudoanonymized mobile data can be of tremendous help (Oliver et al., 2015).

In this section I summarize a project on this topic that we carried out in collaboration with UN Global Pulse and the Polytechnic University of Madrid

(UPM) and with the participation of experts from the Mexican Government and the World Food Program. This summary was published in Spanish in the blog by the Inter-American Development Bank (IDB) (Pastor-Escuredo et al., 2016) and a detailed description of the project and its results can be found in Pastor-Escuredo et al. (2014).

Between 2007 and 2011, the Mexican state of Tabasco suffered numerous floods which caused damage and losses with an economic cost of over 57 million Mexican pesos, according to United Nations Development Programme (UNDP). In this project, we studied the floods that took place at the end of 2009. The goal of the project was to shed light on the value of Big Mobile Data to develop early warning mechanisms and to support decision-making processes when facing a flood. The results of the project showed that it is possible to quickly identify the most affected regions by analyzing the patterns of activity in mobile network towers (Base Transmitter Stations or BTS).

2.2.1 The Value of Data from Multiple Sources to Tackle a Real-life Problem

In this study we measured the social reaction to the floods (measured through the levels of activity in the mobile network towers) as a response to the institutional interventions (namely, the civil protection warnings to the population) and the rain patterns (external factor). In order to measure the social reaction to the floods, we analyzed anonymized and aggregated[1] mobile data, ensuring that privacy was always preserved. The institutional interventions were captured through a search of news, public records and civil protection records about the 2009 and 2010 floods in Mexico. Finally, the rain patterns were measured using open satellite data from NASA.

We used official census data to verify that the geographic distribution of our mobile phone sample was representative of the population in the region. We analyzed aggregated and anonymized mobile data at the cell tower level, in such a way that the cell towers became our sensors of human activity and mobility. By analyzing the patterns of activity in the cell towers before, during and after the floods took place, we were able to automatically detect anomalous behavior which was linked to the areas affected by the intense rains. LANDSAT image analysis enable us to automatically detect the flooded areas in detail. In addition, data from the TRMM-NASA Project provided us with

[1] Aggregated data is data where several measurements are combined. In our case, we combined data from hundreds to thousands of individuals to preserve privacy.

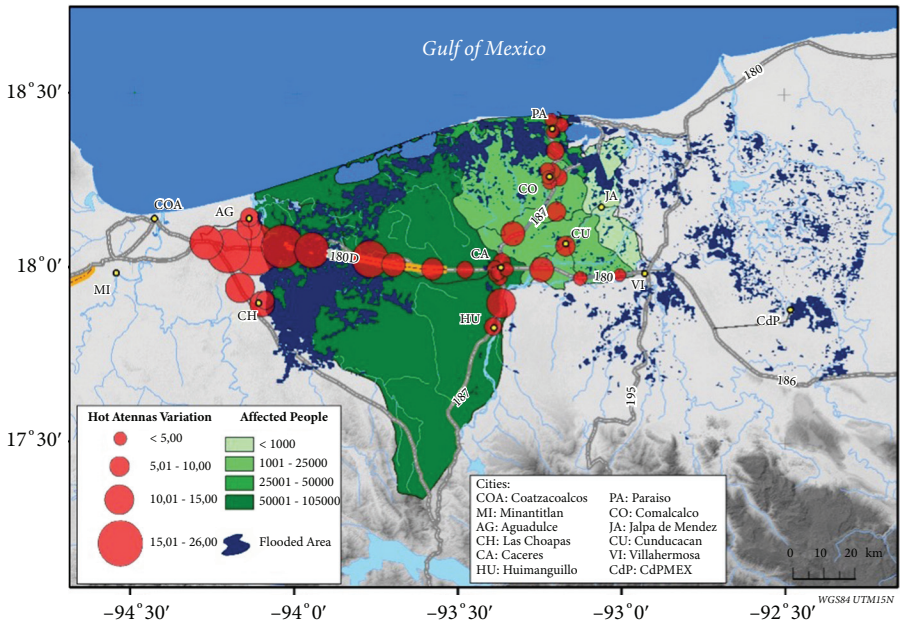


Figure 2.1 Map of the impact of the floods in 2009 in Tabasco, Mexico. The map shows the worst day with the highest values in variation in the activity levels in the cell towers.

Reproduced with permission from Pastor-Escuredo, D., Morales-Guzmán, A., Torres-Fernández, Y., Bauer, J.-M., Wadhwa, A., Castro-Correa, C., Romanoff, L., Lee, J. G., Rutherford, A., Frias-Martinez, V., Oliver, N., Frias-Martinez, E., & Luengo-Oroz, M. (2014). *Flooding through the lens of mobile phone activity*. IEEE Global Humanitarian Technology Conference GHTC'14.

daily information of the overall precipitation which enabled us to compare the evolution in the floods with the patterns of activity in the cell towers. We aligned the different sources of data such that we could quantify the impact of the floods (see Figure 2.1).

2.2.2 Data-driven Conclusions

According to the records, Civil Protection sent an alert signal on the day of maximum intensity of rain. However, the calling patterns observed in the cell towers did not reveal any significant behavioral changes due to this alert. Interestingly, the most significant changes in calling patterns took place in very specific areas that had been severely affected by the rain—such as roads—on the days following the Civil Protection alert. This finding is important because it illustrates how, in the case of floods, the level of citizen awareness might not take place on time, particularly given that institutional alerts seem to be ignored. Given the results of our analysis, we observe that

the patterns of activity in the cell towers are helpful to (1) improve government interventions by automatically and quickly identifying the most affected areas; and (2) to redesign and improve government strategies to alert the population.

2.2.3 Implications of this Study

This use case illustrates the potential of leveraging large-scale, aggregated, and anonymized data to improve our understanding of the impact of natural disasters and to increase the effectiveness of government interventions. The analysis of the data helps improve the effectiveness of the actions to be taken in any of the phases of a natural disaster: (1) during the natural disaster, by enabling automatic detection of the most affected areas which is of paramount importance to optimize the humanitarian intervention and the diffusion of information; (2) after the government interventions have been implemented, by making it possible to monitor the effectiveness of such interventions and to design recommendations to maximize their positive impact; and (3) retrospectively, by analyzing post-hoc data to improve the policies used to prepare, mitigate and recover from a natural disaster.

In the next section, I summarize another example of the value of Big Data to improve the quality of life, specifically regarding criminality in cities.

2.3 Big Mobile Data and Crime

Crime has significant impact on the quality of life and the economic development of a region (Cullen & Levitt, 1999). However, until 2013 crime had not been extensively studied within the area of Big Data for Social Good (Ferrara et al., 2014; Traunmueller, Quattrone, & Capra, 2014; Wang et al., 2013; Catlett et al., 2019; Aghababaei & Makrehchi, 2016; Kadar & Pletikosa, 2018).

Architects and urban planners have studied the relationship between urban dynamics, the specific characteristics of a city and crime. In the 1960s, the social activist Jane Jacobs (1969) emphasized the role of what she referred to as natural surveillance as a key element to reduce crime: as people move within an area, they act as eyes on the street, able to observe what happens around them and hence to protect each other. Jacobs suggested that high diversity in the population and a large number of visitors would therefore contribute to increased security of neighborhoods, leading to less crime.

Conversely, a few years later Newman proposed that high levels of diversity in a population would lead to anonymity which is a key necessary factor for crime (1973). Hence, according to Newman, low levels of population diversity, a low number of visitors and a large percentage of residents are important characteristics to ensure the safety of a neighborhood. Several studies have tried to find evidence to corroborate one or other of these two opposing theories. In the late 1990s, Felson and Clarke (1998) proposed the Theory of Routine Activity which studies how different types of human activities and lifestyles impact the opportunities for crime to exist. One of their findings is that bars and pubs tend to attract crime.

In recent years, criminologists have started to focus their attention on the fact that crime tends to concentrate in certain geographic areas, leading to the concept of crime hotspots. In 2008, criminologist David Weisburd (2008) proposed shifting the paradigm of crime from being centered on people to a paradigm of crime that is centered on the physical space.

Based on this previous work, we framed the challenge of automatically detecting crime through the analysis of Big Data as a place-centric problem. We investigated the predictive value of human dynamics—as inferred through the analysis of large-scale mobile network data and demographic information—to determine if a specific geographic area is likely to be a crime hotspot or not. A complete description of this project can be found in Bogomolov et al. (2014).

2.3.1 Data

The project described in this section leveraged data that was shared during a public Datathon—the Datathon for Social Good—organized by Telefónica, the Open Data Institute and the MIT Media Lab in September of 2013. Among other datasets, participants in the Datathon had access to:

1. Aggregated and anonymized data about human activity and demographics. We refer to this data as SmartSteps data given it was provided by a Telefónica product called SmartSteps (Telefónica UK, 2019), as depicted in Figure 2.2. The spatial granularity of this data is given by square cells, which result from dividing the space in a grid into squares whose size is inversely proportional to the density of cell towers in a specific region, i.e., the higher the density of cell towers, the smaller the size of the grid. For each cell in the grid, the participants in the Datathon had access to several variables computed each hour over a period of

three weeks from December 9–15, 2012 and from December 23, 2013 to January 5, 2014. In particular, the variables include: (1) the estimated number of individuals in each cell computed by extrapolating to the general population the levels of activity in the mobile towers; (2) the estimated gender, age and percentage of people for whom each cell is their home, their work or they are just visitors. Gender and age were provided by the market research firm GFK. Age was given the following bins: 0–20 years, 21–30 years, 31–40 years, etc.

2. Open and geo-localized crime data. This dataset includes all the geolocalized crimes reported in the United Kingdom with a monthly temporal granularity. The data shared in the Datathon corresponded to the crimes from December 2012 and January 2013. We used the data from December to train the models, and the data from January to test it. The spatial granularity of this data is the Lower Layer Super Output Area (LLSOA) which is the census sector used in the United Kingdom. The crimes are classified into 11 types (e.g., anti-social behavior, robbery, violent crime, etc.).
3. Census data. The official London census data consists of 68 variables which characterize the population in each LLSOA, including the number of households, the employment percentage, the proportion of people older than 65 years, the percentage of immigrants, the ethnicity, the mother language, average income, average house price, number of cars per household, percentage of children in families without a job, happiness levels, life expectancy, etc.

2.3.2 Crime Model

We formulated our problem as a binary classification problem applied to each cell in space. Given that the crime dataset is unbalanced, we used the median value to define if a cell is a crime hotspot or not: we assigned a 0 to all the cells whose number of monthly reported crimes was equal or less than the median (=5 in our case) and we assigned a 1 if the monthly number of reported crimes was greater than the median, and hence a crime hotspot.

We used 80% of the cells to train the model and 20% of the cells to test it. A fundamental step requires geo-referencing all the data using the same geographic reference system. The SmartSteps cells are square, while the LLSOA have arbitrary shapes. In order to unify both representations, we assigned the criminal and census data to the SmartSteps cell with the closest centroid to the LLSOA.

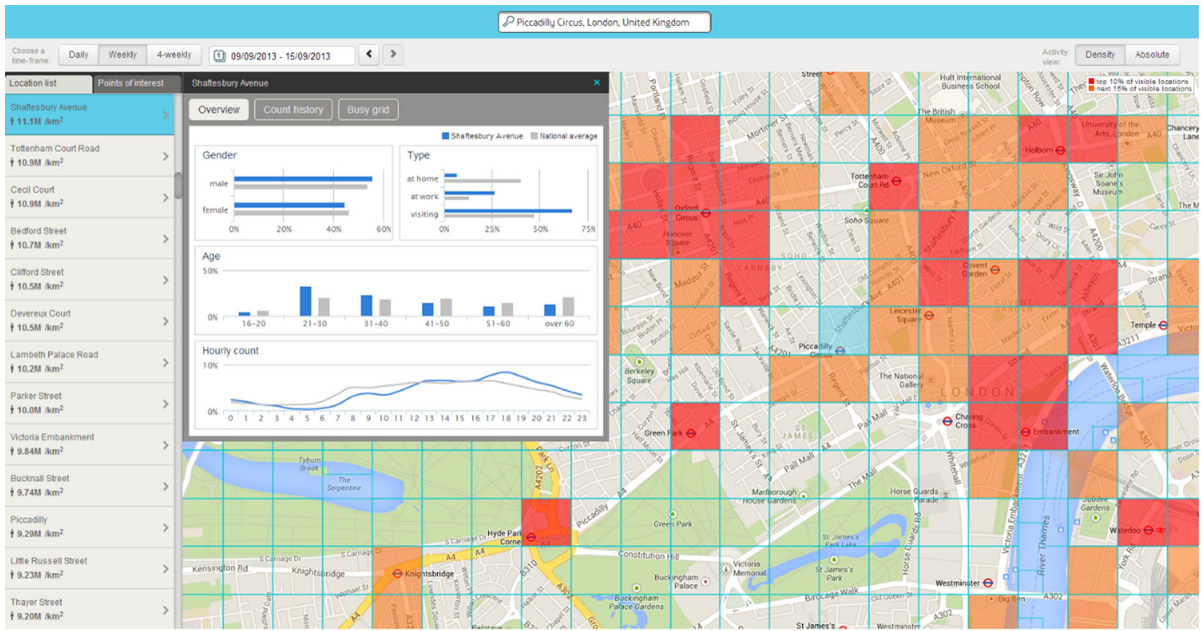


Figure 2.2 Visualization example of the SmartSteps product. Spatio-temporal information about the dynamics of the city can be obtained by combining aggregated and anonymized mobile data with demographic data. This information is important for our problem at hand, namely crime prediction.

Reproduced with permission from Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile phone data. *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 2014)*, 427–434.

Previous studies have found that variables that measure the diversity and regularity of human behavior are important for its characterization. In particular, the entropy of the variables has been used to successfully estimate the socioeconomic levels or regions and cities (Eagle et al., 2009), mobility and consumption patterns. Hence, in our use case we computed the mean, the median, the standard deviation and the minimum and maximum values of the Shannon entropy of the Smart-Steps data. In addition, we computed all previous statistics (i.e., mean, median, etc.) on temporal windows of different lengths, namely one hour, four hours and one day.

In the case of the census, we used the 68 original variables without transformations.

Given that we had 68 census variables, we applied a state-of-the-art feature selection method (namely, the mean decrease in Gini coefficient) to select the 68 most predictive variables from the SmartSteps dataset. As a result, we had two different models of similar dimensionality: one built using the Smart-Steps variables (model based on population dynamics) and the other model built using the census variables (model based on traditional variables).

Regarding the classification algorithm, we used Random Forests (Breiman, 1999) because they satisfy the maximum margin property, they do not require parameter tuning, and more importantly they do not require the specification of the feature vector space, as other models require, including Support Vector Machines.

We evaluated and compared three models, in addition to a baseline model based on a majority classifier which would return the majority class for each SmartSteps cell. The three compared models are: (1) a model that uses the census variables as input; (2) a model that uses the SmartSteps variables as input; and (3) a joint model that takes both census and SmartSteps variables as input.

In terms of performance, we computed accuracy (Acc.), the F1 measure (the harmonic mean between precision and recall) and the area under the ROC curve (AUC). Table 2.1 summarizes the results of the model evaluation. Figure 2.3 depicts the classification results of the different cells in London's metropolitan area as crime hotspots or not.

As shown on the Table 2.1, the models built using features about population dynamics—alone and in combination with census features—(SmartSteps and SmartSteps+Census) outperform both the model built using traditional census features and the baseline model.

Next we analyze the performance of the models for the class that corresponds to a high number of crimes (value 1 in the model or positive class). We look at the true positive rate—that is, the proportion of crime hotspot

Table 2.1 Performance comparison of different models.

Model	Acc., %	Acc. CI, 95%	F1, %	AUC
Baseline Majority Classifier	53.15	(0.53, 0.53)	0	0.50
Borough Profiles Model (BPM)	62.18	(0.61, 0.64)	57.52	0.58
SmartSteps	68.37	**(0.67, 0.70)**	65.43	0.63
SmartSteps + BPM	**69.54**	**(0.68, 0.71)**	**67.23**	**0.64**

Source: Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile phone data. Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 2014), 427–434.

cells that have been correctly classified by the model. The models that use SmartSteps and SmartSteps+Census data obtained good true positive rates of 74.2% and 73.9%, respectively. Conversely, the model that uses Census data only achieves a 68.81% true positive rate, which is more than 5% below the models that include SmartSteps data.

Regarding the true negative rate—that is the proportion of cells that are not crime hotspots and have been correctly classified as such by the model, all the models obtain worse results: 63.07% for the SmartSteps+Census model, 61.06% for the SmartSteps model, and 54.66% for the Census model. Note that the SmartSteps model has a true positive rate that is roughly 10% better than the true negative rate. In other words, the models are better at detecting the cells that are crime hotspots than the cells that have low levels of crime. In our use case, it is more important to have good performance in the positive class given that from a social perspective it is less dangerous to erroneously consider a cell to be a crime hotspot when it is not, than the other way around. As shown in the results, the proposed model brings significant advantages when compared to the state-of-the-art models (baseline and BPM). The subplots in Figure 2.3 illustrate the performance of the model (right image) when compared to the ground truth (left image).

2.3.3 Implications

The results presented in the previous section show that data about aggregate human dynamics and basic demographic data with a daily or monthly time window bring significant value to automatically detect crime hotspots, yielding models that outperform models built using census data (demographics, ethnicity, employment, etc.). Census data is typically very detailed, but the data is very expensive to obtain and hence has very large temporal granularity

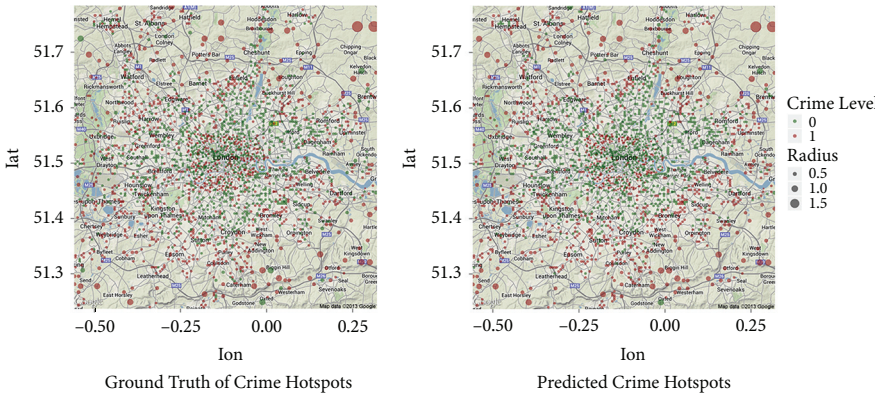


Figure 2.3 Ground truth of crime hotspots and predicted crime hotspots.

Reproduced with permission from Bogomolov, A., Lepri, B., Staiano, J., Letouzé, E., Oliver, N., Pianesi, F., & Pentland, A. (2015). Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. *Big Data*, *3*(3), 148–158. https://doi.org/10.1089/big.2014.0054

(a census is typically built only every 10–12 years). Given the low frequency of this type of data and the highly dynamic nature of cities, its value to automatically detect crime hotspots is limited in our experience. Conversely, the data obtained from the mobile network infrastructure—combined with basic demographic information (gender and age)—has higher predictive power for this task, despite being less detailed than the census. It seems that (spatially and temporally) fine grained information about urban dynamics is important to model crime.

When we look at the 20 most predictive variables in the models, we observe:

1. In the joint model (SmartSteps+Census), there is no census variable among the top 20 most predictive variables;
2. Variables that are computed with a daily time window have more predictive power than variables computed with a monthly time window;
3. Variables about the percentage of people for whom a certain cell is their home—both daily and monthly—seem to be very important. In fact, 11 of the 20 most predictive variables are related to the home concept. Newman's "defensible space theory" postulates the relevance of variables related to the number of residents in a place to characterize areas with low crime. However, in contrast with Newman's theory, we find positive correlations between the home variables and crime. Hence, our experiments do not corroborate Newman's theory. Recent work by Traunmueller et al. (2014) has found similar results.

4. The entropy of the variables also plays an important role, as 8 of the 20 most predictive variables are entropy variables. Entropy captures the predictable structure of a variable. In our case, it would correspond to the predictable elements of a place regarding the types of people that spend time in such a place throughout the day. A place with high entropy would have a lot of variety in terms of the kinds of people who visit it daily; conversely, a place with low entropy would be characterized by a high regularity in the people who visit it daily. Both our results and the results reported by Traunmueller et al. (2014) corroborate Jacobs' natural surveillance theory: a high diversity of people in a place (both in terms of demographics and residence status) enables the existence of "eyes on the street" that monitor the environment and reduce crime rates.

The approach proposed in this section could have practical implications for police departments and city government officials, providing them with guidelines regarding where to invest efforts and how to best respond to crime.

2.4 Conclusion

The potential of Big Data to improve the world is immense. Even though we are at the beginning of this revolution, it is a necessary and important revolution. Through the analysis of Big Data using advanced analytics and machine learning techniques, we have the opportunity to be able to make better, fairer and more timely decisions which are evidence-based. The two use cases described in this chapter illustrate the potential of leveraging Big Data to tackle pressing societal challenges in countries with different socioeconomic levels and infrastructures.

However, we still need to overcome a diverse set of challenges before we can effectively realize such potential in a sustainable way, including:

- *Regulatory challenges*, given that existing regulation is typically not updated to include use cases such as the ones described in this chapter. In addition, there are no clear principles regarding how to securely share aggregated and anonymized data for humanitarian purposes. In this direction, I would highlight the OPAL[2] project, a techno-social platform of open algorithms that can be securely run within the premises of the owners or custodians of the data, in order to compute key indicators

[2] Please see: https://www.opalproject.org/.

and the report published in 2021 as a result of the work by the high level expert group at the European Commission on "Business to Government Data Sharing" (European Commission, 2020).

- *Social challenges*. It is of paramount importance to take into consideration non-intended negative consequences that could result from the data analysis, even when the goal of such an analysis is to have positive social impact. In addition, it is key to be aware of the risk of creating—or increasing—a divide between those who have access to the data and to the skills needed to extract valuable insights from it, and those who do not have it. Therefore, if we want to realize the potential of Big Data for Social Good, we need to invest in data literacy programs;
- *Technical challenges*, of diverse nature. This is an emergent research area with many research challenges still to be solved. We need to understand how representative the data is, which potential biases are present in the data and to which degree we can generalize the conclusions obtained through the analysis of the data. In order to tackle real problems, we need to combine data from different sources, which poses its own challenges. Real-time analysis is critical in some use cases (such as natural disasters). However, we are still far from achieving this goal as most available data is post hoc or off-line. In many use cases we do not have reliable ground truth such that it is difficult—if not impossible—to validate and quantify the impact of the results without carrying out interventions derived from the data analysis. Moreover, successfully carrying out interventions in the real world is complex. The transparency and interpretability of the data-driven algorithms that are used to make decisions for public good is also key (Lepri et al., 2017). Finally, most of the up-to-date published research focuses on finding significant correlations between the variables of interest. However, there are important opportunities that arise when we can attribute a cause-and-effect relationship to the variables, which is difficult to do.
- *Privacy and security challenges*. Privacy risks should be minimized and quantified (de Montjoye et al., 2013). Transparency, access control, bias minimization, security and traceability of the data are of paramount importance (Lepri et al., 2017). It is also necessary to define and act according to a clear code of conduct and ethical principles when it comes to data handling and analysis.
- *Way of working challenges*. Big Data for Social Good use cases are by definition multi-disciplinary and multi-institutional, because they require joint work by experts in different fields (e.g., data science, public health, emergency relief, etc.). In order to ensure that these types of projects become a reality, multi-disciplinary teams from different

institutions need to effectively work together. These projects typically require signing collaboration agreements and NDAs between different institutions (e.g., academic, companies, governments, NGOs, etc.) which could delay making progress on the projects.

- *Financial challenges*. Big Data for Social Good projects do not necessarily need to be undertaken for free or carried out on a pro-bono basis. In fact, if we are not able to identify sustainable business models for these types of projects, they will probably not prosper beyond the pilot phase. Hence, it is of utmost importance to define financially sustainable models of bringing these projects to production.

Fortunately, none of these challenges is unsurmountable. The opportunity of using Big Data for Social Good is too large and too necessary not to join forces and do our best to address the challenges above. Thus, in order to fully realize the potential of these new data sources, we need to develop new private-to-public data sharing models that are legally compliant, socially acceptable, technical feasible, commercially and financially viable and sustainable over time. Real-time access to relevant data is particularly important in some use cases, such as when facing a natural disaster, which adds another element of complexity.

When we think of Big Data, I encourage you to not only think of it as a business opportunity, but also to consider its value to have positive social impact. Moreover, having access to real-time measurements is of utmost importance to better tackle numerous emergency situations, such as natural disasters as real-time data enables a fast response that is adjusted to the actual situation with the right allocation of resources and enhanced situational awareness to support evidence-driven decision-making. It is without a doubt this potential what motivates me to work and carry out research in this fascinating field.

2.5 Co-authors

The projects described in this chapter are the result of collaborations with the following co-authors of the work: Dr. Vanessa Frias-Martinez, Dr. Enrique Frias-Martinez, Dr. Alex Pentland, Dr. Bruno Lepri, Dr. Jacopo Staiano, Dr. Emmanuel Letouzé, Dr. Andrey Bogomolov, Dr. Miguel Luengo-Oroz, Dr. Alex Rutherford, Jong Gun Lee, Liudmyla Romanoff, Carlos Castro-Correa, Amit Wadhwa, Jean-Martin Bauer, Dr. Yolanda Torres-Fernandez, Dr. Alfredo Morales-Guzman, Dr. David Pastor-Escuredo, and Dr. Fabio Pianesi.

References

Accenture. (2017). *Insight big data research*.

Aghababaei, S., & Makrehchi, M. (2016). Mining Social Media Content for Crime Prediction. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 526–531. https://doi.org/10.1109/WI.2016.0089

AON. (2023). *Weather, Climate and Catastrophe Insight 2022*. Aon plc. https://www.aon.com/getmedia/f34ec133-3175-406c-9e0b-25cea768c5cf/20230125-weather-climate-catastrophe-insight.pdf

Bayoumi, T., Quayyum, S. N., & Das, S. (2021). Growth at risk from natural disasters. *IMF Working Papers*, *2021*(234). https://doi.org/10.5089/9781513597652.001.A001

Bogomolov, A., Lepri, B., Staiano, J., Letouzé, E., Oliver, N., Pianesi, F., & Pentland, A. (2015). Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. *Big Data*, *3*(3), 148–158. https://doi.org/10.1089/big.2014.0054

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. *(2014)*. Once upon a crime: Towards crime prediction from demographics and mobile phone data. *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 2014, 427–434.*

Breiman, L. (1999). *Random Forests: Random Features* (Technical Report 567). Department of Statistics, UC Berkeley.

Catlett, C., Cesario, E., Talia, D., & Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive and Mobile Computing*, *53*, 62–74. https://doi.org/10.1016/j.pmcj.2019.01.003

Cullen, J. B., & Levitt, S. D. (1999). Crime, urban flight, and the consequences for cities. *The Review of Economics and Statistics*, *81*(2), 159–169. https://doi.org/10.1162/003465399558030

DataIQ. (2017). *Networking data scientists for social good*.

Data-Pop Alliance. (2015). *Data-Pop Alliance*. https://datapopalliance.org/

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*(1), Article 1. https://doi.org/10.1038/srep01376

Eagle, N., Pentland, A. (Sandy), & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278. https://doi.org/10.1073/pnas.0900282106

European Commission. (2020). *Towards a European Strategy on Business-to-Government Data Sharing for the Public Interest: Final Report Prepared by the High Level Expert Group on Business to Government Data Sharing*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/731415

European Commission. (2021, May 9). *Results of the new European Data Market study 2021–2023*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/results-new-european-data-market-study-2021-2023

Felson, M., & Clarke, R. V. (1998). *Opportunity Makes the Thief: Practical Theory for Crime Prevention* (98; Police Research Series Paper). Home Office.

Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). *Detecting criminal organizations in mobile phone networks*. https://doi.org/10.1016/j.eswa.2014.03.024

The Flowminder Foundation. (2023). *Flowminder—Innovative solutions from mobile data analytics*. https://www.flowminder.org

Fortune Business Insights. (2022). *Big Data Analytics Market Forecast, 2022–2029*. Fortune Business Insights.

Frías-Martínez, V., Rubio, A., & Frías-Martínez, E. (2012). Measuring the impact of epidemic alerts on human mobility using cell-phone network data. *PURBA*. https://www.semanticscholar.org/paper/Measuring-the-impact-of-epidemic-alerts-on-human-Fr%C3%ADas-Mart%C3%ADnez-Rubio/7d1b5577e6c078f9d148a913c07fd77a78dc5bb2

GermanWatch. (2020). *Global climate risk index 2020*.

GSMA. (2017). *Big data for social good*.

Jacobs, J. (1969). *The Death and Life of Great American Cities*. Modern Library.

Kadar, C., & Pletikosa, I. (2018). Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science*, *7*(1), 1–27, Article 1. https://doi.org/10.1140/epjds/s13688-018-0150-z

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good. In T. Cerquitelli, D. Quercia, & F. Pasquale (Eds), *Transparent Data Mining for Big and Small Data* (pp. 3–24). Springer International Publishing. https://doi.org/10.1007/978-3-319-54024-5_1

Letouzé, E. (2012). *Big Data for Development: Opportunities and Challenges*. UN Global Pulse. https://www.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay2012.pdf

Mancini, F., & O'Reilly, M. (2013). New technology and the prevention of violence and conflict. *Stability: International Journal of Security & Development*, *2*(3), https://doi.org/10.5334/sta.cp

Newman, O. (1973). *Defensible Space: People and Design in the Violent City*. Architectural Press.

Oliver, N., Matic, A., & Frias-Martinez, E. (2015). Mobile network data for public health: Opportunities and challenges. *Frontiers in Public Health*, *3*, 189. https://doi.org/10.3389/fpubh.2015.00189

Open Data Institute, Telefonica, Campus Party, & MIT. (2013). *Datathon for social good*.

Orange. (2015). *Data for development challenge senegal: Book of abstracts*.

Pastor-Escuredo, D., Morales-Guzmán, A., Torres-Fernández, Y., Bauer, J.-M., Wadhwa, A., Castro-Correa, C., Romanoff, L., Lee, J. G., Rutherford, A., Frias-Martinez, V., Oliver, N., Frias-Martinez, E., & Luengo-Oroz, M. (2014). Flooding through the lens of mobile phone activity. *IEEE Global Humanitarian Technology Conference GHTC'14*.

Pastor-Escuredo, D., Morales, A., Torres, Y., Frias Martinez, V., Frias Martinez, E., Oliver, N., & Luengo-Oroz, M. (2016, February 25). Cómo el Big Data ayuda en la Gestión de Desastres Naturales: El caso de Tabasco en 2009. Abierto al Público. https://blogs.iadb.org/conocimiento-abierto/es/como-el-big-data-ayuda-en-la-gestion-de-desastres-naturales-el-caso-de-tabasco-en-2009/

Ricci, F., Rokach, L., Shapira, B., & Kantor, B. P. (2010). *Recommender Systems Handbook* (1st edition). Springer-Verlag.

Salah, A. A., Pentland, A., Lepri, B., Letouze, E., Vinck, P., de Montjoye, Y.-A., Dong, X., & Dagdelen, O. (2018). *Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey* (abs/1807.00523). CoRR.

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of Socioeconomic Levels Using Cell Phone Records. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds), *User Modeling, Adaption and Personalization* (pp. 377–388). Springer. https://doi.org/10.1007/978-3-642-22362-4_35

Taylor, P. (2023, November 16). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Telecom Italia. (2014). *Big data challenge*.

Telefónica UK. (2019). *Smartsteps*. https://tis.smartsteps.telefonica.com/#/

Traunmueller, M., Quattrone, G., & Capra, L. (2014). Mining mobile phone data to investigate urban crime theories at scale. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11–13, 2014. Proceedings 6* (pp. 396–411). Springer International Publishing.

United Nations. (2017). *The Global Partnership for Sustainable Development Data*.

Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). Learning to detect patterns of crime. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23–27, 2013, Proceedings, Part III 13* (pp. 515–530). Springer Berlin Heidelberg.

Weisburd, D. (2008). *Place-based policing* (9; Ideas in American Policing, pp. 1–16).

3

Digital Forms of Exclusion, Big Data, and Measurement of UN Sustainable Development Goals in Africa

An Ecological Systems Theory Perspective

Bitange Ndemo

3.1 Introduction

The most profound indicator of social change is the capacity of individuals to leverage verifiable methods, such as data, to enhance their communities and society. Bronfenbrenner (1977, 1979) conceptualizes this evolution of identity as "development in context," rooted in human ecology theory. African nations have long pursued the realization of this theory, and the current African data revolution offers an unprecedented opportunity for African countries to take decisive steps towards securing their autonomy. Studies (Ndemo & Mkalama, 2024) indicate that data and emerging technologies, including artificial intelligence, are pivotal in the successful elimination of poverty, aligning with the first Sustainable Development Goal.

The concept of "Big Data" was recognized academically in February 2011 when Kenneth Cukier published a special report in *The Economist* entitled "Data, Data Everywhere." Manyika et al. (2011), from the McKinsey Global Institute, later published "Big Data: The Next Frontier for Innovation, Competition, and Productivity." Since then data has become a tool for decision making.

A bottom-up process of encouraging communities to embrace data in decision-making is under way for the African continent as it works toward meeting the United Nations (UN) General Assembly Sustainable Development Goals (SGDs). The UN Economic Commission for Africa launched the African SDG progress online evidence-driven policymaking tool based on

Bitange Ndemo, *Digital Forms of Exclusion, Big Data, and Measurement of UN Sustainable Development Goals in Africa*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Bitange Ndemo (2025). DOI: 10.1093/oso/9780198879145.003.0004

Agenda 2063, "The Africa We Want," which includes meeting the SDGs. The platform includes information on SDG indicators at regional, subregional, and national levels, alongside tracking and probability estimates based on the 2030 goals (UNECA, 2022).

When the SDGs were launched, it was expected that participating countries would intuitively track their progress according to human ecology theory predictions, in which Huntington (2000) suggested that humans have the propensity to pursue intuitive knowledge, intelligent relationships, mutual understanding, and an ability to collaborate about experiences and the natural world. Ostensibly, this type of rapport will build toward economically sustainable development (Choi et al., 2015). Human ecology theory borrows from the ecological economics of de Groot et al. (2002), who adopted organic methods of process valuation to address the inequities between market growth and biodiversity loss. For example, the implications of poverty eradication at the expense of environmental loss suggest that the net progress of green development may be zero. Leveraging data to study the advancements made toward achieving SDGs should therefore clearly illuminate ways to balance development and environmental sustainability.

Although the SDGs were adopted in 2015 based on the assumption that participating nations would naturally leverage data-driven decision-making processes, governments, development institutions, and other nongovernmental organizations (NGOs) have struggled to make them work, especially in Africa. The SDG goals and their indicators include metrics from the earlier Millennium Development Goals (MDGs), which included eradicating poverty and hunger, improving human health, conserving water, cleaning the environment, industrializing intelligently, and benefiting urbanization. Crucial to their success, strong governance was assumed. However, a motivated data-centric "digitized" culture is also needed alongside the statistical systems that are to be used to measure and incentivize progress (SDSN, 2015).

Prior to 2015, governments and regional affiliates developed moderately data-based consensus judgments in the lead-up to the creation of the SDGs. For example, the African Data Consensus (2015) originally advocated for a persistent data revolution to fuel social, economic, and structural change across all of Africa. The envisioned revolution was to enable African countries to track their progress in meeting their national agendas.

Orthodox sustainable development is a normative concept that calls for global societies to work together to balance economic, social, and environmental objectives. When countries pursue gross domestic product (GDP)-focused objectives, it thwarts greater social and environmental objectives. Notably, the UN SDGs are designed to guide countries in achieving

economic, social, and environmental objectives in harmony with national goals, thereby leading to greater levels of wellbeing for present and future generations (Helliwell et al., 2015).

Even as the world began implementing these types of goals while integrating big data as the key mode of their achievement, heavy skepticism and reluctance persist. New research (i.e., Mayer-Schönberger & Cukier, 2013) suggests that the focus on SDGs may have unintentionally marginalized millions of people by virtue of their inability to participate in the "digital revolution." If the big data construct is to epistemologically shape government and corporate decision-making processes in response to SDGs, then much more attention is needed on the source and nature of the data. Notably, experts have shown a startling lack of regard for the accuracy of information and the influence of bias when contemplating big data-driven processes.

With big data, the lack of structure (i.e., its "messiness") is considered a virtue (Mayer-Schönberger & Cukier, 2013) because their vastness and volume offer powerful complex correlations that can lead to incredibly powerful conclusions that would not have been intuitive to humans alone. Unfortunately, this analytical power has overshadowed the need to discern the sources of data and their reliability (Lerman, 2013).

It is well known that big data collection, manipulation, storage, and interpretation require huge investments in technological and human resources. As such, this chapter answers the following research questions:

- Is Africa investing sufficient resources to meet big data demands for SDG accomplishment?
- To what extent are big data being used to create inclusive policies?

By answering these questions, this study highlights the utility of big data in terms of an African strategy for mitigating the many social problems prevalent in the region.

3.2 Theory of Human Ecology

Human ecology theory dates back to 1887, when Ellen Swallow Richards conducted pioneering research on drinking water in Massachusetts, which led to improvements in water quality and modern sewage treatment. Richards stated, the environment in which people live is the environment to which they react and which they maintain. If the environment is healthy, great; if not, life

there will be of bad quality. This is a systems theory method where inputs trigger actions that trigger procedures that trigger results that trigger feedback. Significantly, this philosophy applies to big data because the outcomes are to be used for social and industrial progress, and if the quality of life is low, "where the data [are] good, so be it." The knowledge gained from the data is expected to inform new intervention decisions, such as reducing carbon emissions to improve environmental sustainability outcomes or inclusionary policy implementations.

Environmental sustainability is a critically important social problem, and although Richards' study was carried out in the late nineteenth century, it resonates with today's environmental discourse, given the fact that the author foresaw ecology as "a means for applying principles, methods, and results of science to improve people's lives and their environment." In contrast, general systems theory (Bertalanffy, 1968) focuses on iterative changes with respect to entities and their relationships that are agnostic of environmental impacts. Unfortunately, this theory is the most often applied to societal development decisions. On the other hand, the ecological systems theory developed by Bronfenbrenner improves both Richards' and Bertalanffy's works and is now one of the most widely recognized explanations of how social settings can affect human development.

This article argues that any ecological systems position must recognize both ordered and unordered data on people and their interactions within an environment as crucial to formulating and achieving beneficial goals. Moreover, ecological systems reflect sequences of biological transitions that require environmental sustenance and survival capabilities. As such, they are beneficial to survival (e.g., modeling healthcare, family, income, career, and educational solutions). An epistemology based on these perspectives would enable practitioners to better assess factors relevant to vital social problems (Hepworth et al., 2010).

3.3 Emerging Forms of Exclusion Resulting from a Lack of Enabling Technologies and Data

There are several contexts in which exclusionary data-sharing practices emerge. Recently, several scholars (e.g., Mayer-Schönberger & Cukier, 2013 and Seifert et al., 2021) noted that vulnerable groups (e.g., senior citizens and children) often fail to utilize digital services because they are reluctant or unable to use the Internet, or they lack the technical expertise to submit queries and interpret results. Moreover, if not carefully

designed, the data policies themselves can become barriers to a large portion of the larger population. Foremost, many valuable digital data sources are not shared publicly. Global mobile telecommunication systems research (GSMA, 2014) has shown that there are huge disparities between the number of individual mobile subscribers (people) and the number of mobile connections (resources). That is, multiple smartphones are owned by single individuals, implying that non-smartphone owners are generally excluded from statistical evaluations. Hence, extant measurements of mobile market penetration are often misleading, and there remains a vast population that lacks the tools needed to participate in a data-driven culture.

3.4 Big Data Exclusions at the Individual Level

Let us consider two hypothetical people: a 25-year-old male from Nairobi and a 30-year-old female from rural. Like his millennial counterparts, the urban Nairobi male owns a smartphone, has multiple social media accounts, accesses broadband, rarely carries cash, and travels extensively. The female counterpart from a rural county in Kenya lives near Amboseli National Park, which is sparsely populated. Telecommunication companies avoid investing heavily in these areas owing to their low returns on investment. Moreover, because the rural individual owns a simple-feature phone (i.e., cheap and austere), she cannot connect to broadband networks without traveling 60 km to the nearest large town. She lives in poverty and teaches at a nearby private nursery that pays her cash. Hence, she remains disconnected from the social infrastructure and rarely travels outside of her county. Notably, she also lacks online social networks and remote financial connections.

People like these two examples are found in all nations, but the second example describes types that are most often found in developing countries. Despite the ubiquity of smartphones, a significant number of people remain underserved. Hence, critical data sources and their requisite applications are inaccessible to the people who need them. Jackson et al. (2008) suggested that this digital divide is largely congruent with the race and gender divides in advanced countries. However, in developing countries, Ahmed (2007) warned that the digital divide is a clear economic division, as global economies increasingly rely on data-driven capabilities. Hence, poor people do not receive commodities. Notably, heavy infrastructure investments are needed to support these growing economies and emerging democracies, without which, we may be headed toward a global depression.

3.5 Failure to Exploit Data Resources Leads to Institutional Exclusions

Institutions worldwide weigh the returns on investment of investing in massive big data implementations (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011), yet very few African institutions have demonstrated strident commitments. Odhiambo and Umar (2019) reported on this phenomenon in Nigeria, describing its struggle with assessing UN SDGs. A review revealed that only five African universities (mainly in South Africa and Kenya) offered data analytics coursework. Kiron et al. (2011) led a collaborative study with the Massachusetts Institute of Technology's Sloan Management Center and IBM's Institute for Business Value, finding that management's view of the correlation between performance and big data analytics has important implications for the construct's adoption and its potential to influence growth, efficiency, and competitive differentiation (LaValle et al., 2011).

From these studies and others, it is believed that any failure to exploit data resources in support of decision-making processes will lead to corporate and individual exclusions from global markets. Multinational organizations that have already invested in data analytics tend to outperform their top competitors, and their initial investments require a data science culture to be adopted across its internal organizations. In this sense, the pursuit of UN SDGs democratizes this cultural pursuit. Hence, it is not focused on market competition.

3.6 Use of Big data in Measuring SDGs

A report by the Independent Expert Advisory Group (IEAG, 2014) noted that any data revolution that would support sustainable global development would integrate new and traditional forms of data to produce faster, more relevant, detailed, and high-quality decision-level information to illuminate valid courses of action and risks for many complicated situations. Such a revolution would be accompanied by ubiquitous access and transparency, which would empower policymakers to act more reasonably and optimistically while remaining accountable for their decisions (IEAG, 2014).

The IEAG report found that "data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time, designing, monitoring, and evaluating effective policies becomes almost impossible"

(IEAG, 2014: 2). Some countries (e.g., Korea) have developed different measurement models for specific SDGs. Choi et al. (2015) noted that urban information systems can collect, analyze, produce, and manage various data products pertaining to social, economic, and environmental phenomena in urban areas. Hence, they may soon serve as data providers, analytical engines, and information-sharing platforms for the wider provision and assessment of SDG attainment.

Several global NGOs, especially those tackling poverty, have taken steps to assist developing countries with SDGs. The International Water Management Institute (2015) applied a "security" focus to SDG #6 (Clean Water and Sanitation), owing to the need to protect public health. These and other detailed frameworks distinguish the human needs-based SDG approach from the original MDGs.

3.7 Participant–Observer Methodologies

The author of this article spent six years collecting data for this study using a disguised naturalistic observation method. Hence, observations were made as unobtrusively as possible at a series of scholarly big data conferences so that participants would remain unaware that they were being studied. This technique is considered acceptable as long as the participants remain anonymous and the data collection occurs in a public setting (Schmuck, 1997, Kawulich, 2005). Bernard (1994) showed that this type of strategy allows the investigator to act as a "participant–observer," in which anonymous interviews can be conducted without the knowledge of the subject, making it far less obtrusive than more deliberate methods. Moreover, this method strongly reduces the risk of response bias and similar issues. Kawulich (2005) noted that this qualitative method is useful for understanding individual and cultural mechanisms at work.

Using this process and relying on its underlying philosophies, the author observed and documented the perceived emotions, nonverbal expressions, communication methods, and other expressive interactions of participants at these conferences (particularly those of African delegates). Data from all events were later assessed and coded to help the author better understand the nature of African countries' participation and willingness to leverage big data models to pursue and assess their respective SDG efforts. It was of particular interest to observe and assess conference attendee reactions to policymakers as they proposed strategic measures related to big data and SDG accomplishment. One of the most pertinent conclusions is that almost everyone

in attendance agreed that big data would be useful both as a catalyst and an assessment tool. The main reason that this is pertinent is that there has since been a stark lack of action.

3.7.1. Applying an Ecological Systems Perspective to the Use of Big Data and Visualization

This author used the ecological framework of Bronfenbrenner (1977; 1979) to classify the levels of interactions among governmental knowledge centers, from which the required big data networks will be designed. This framework is built upon general systems theory to help illuminate African countries' sensitivities toward environmental protection while considering pragmatic big data integration measures. Three of Bronfenbrenner's ecological system variables were examined. *Interactivity* reflects the behaviors and relationships connecting actors and infrastructure. *Transitions* more directly reflect economic development, and *policy responses* deal with the pragmatic approaches applied to resolving social problems. Notably, the wellbeing and prosperity of the participants are not considered.

All other factors being equal, if a country's policymaking framework and information and communications technology (ICT) infrastructure are integrated, the effects on its citizen's prosperity may be positive overall. Otherwise, the ICTs and their policies will probably become a hindrance.

3.8 Findings: Selected Conferences and Workshops

This section describes Kenya's recent open-data initiatives and their impact on measuring the achievement of SDGs. Much of the information relayed in this section derives from the author's former position as a Permanent Secretary (PS) at the Ministry of Information and Communications in Kenya. As a PS, the author was the chief policymaker representing Kenya at international conferences, which included drafting the Africa Data Consensus for the UN SDGs.

3.8.1 How Kenya Created its Open-data Initiatives

Following the enactment of the 2010 Kenyan Constitution, this author developed policies that reflected citizens' rights to access information. Article 35 of the Constitution states:

> Every citizen has the right of access to—(a) information held by the State; and (b) information held by another person and required for the exercise or protection of any right or fundamental freedom . . . (3) The State shall publish and publicize any important information affecting the nation.

From this article, the author drafted an "Access to Information" bill, but it took 10 years to become an Act of Parliament. The delays were caused primarily by fears (both public and political) that the law would expose questionable personal and organizational activities. However, the Act was finally passed, owing to the strong influence of US President Barak Obama, who has strong Kenyan ancestry. His efforts in passing the US' Open-Data Initiative Act influenced President Kibaki of Kenya to provide support, and it influenced me to work diligently to help garner World Bank backing. Hence, the Kenya Open-Data Initiative (KODI) was born in YEAR against the wishes of many obstinate cabinet members.

Kenya was the first Sub-Saharan country and the second on the African continent after Morocco to freely share public data under law. Soon thereafter on DATE, the UN General Assembly launched the multilateral Open Governance Partnership (OGP) "to promote transparency, empower citizens, fight corruption, and harness new technologies to strengthen governance" (OGP website, https://www.opengovpartnership.org/process/joining-ogp/open-government-declaration/).

3.8.2 UN General Assembly Launch of OGP Initiatives

In 2011, heads of state from Indonesia, Brazil, Norway, Mexico, the Philippines, South Africa, the UK, and the USA validated their OGP and developed action plans with the assistance of prominent leaders of civil organizations. Another 38 governments received extended invitations. Unironically, many of the African government representatives in attendance struggled to understand the implications of big data integration in terms of a civil, collaborative society. The Kenyan delegation strongly facilitated the unity in policy that eventually emerged. However, it could not secure the full investment of African governments.

3.8.3 Brasilia OGP Meeting

At the subsequent 2012 Brasilia OGP meeting (four years prior to the publication of the UN SDGs), participating countries met to select from among 51 renewed MDG themes and develop commitment statements.

Notably, the African delegates were well engaged, and the expected eight participants had surprisingly increased to 50. Despite not having been invited, many additional energetic and skeptical African national delegates attended the meetings, and this author was present to facilitate and explain the reasons behind the new requirements for open data, about which examples from the governments of the USA and UK were highly influential.

Notably, some delegates were legitimately concerned that these efforts might reflect subversive neo-colonialist measures to once again appropriate valuable African resources (i.e., data). One delegate asked if these "shenanigans" were related to the ongoing "WikiLeaks situation." That is, there was a tangible fear of data spillage and theft. Although healthy in most cases, African minsters' skepticism has hindered progress, notwithstanding the excellent policy documents now in place.

3.8.4 African Data Consensus in Addis Ababa

Following the SDG publication in 2015, the African Data Consensus met in March in Addis Ababa, Ethiopia. The sessions were organized and sponsored by the Economic Commission of Africa, African Development Bank, African Union Commission, and the UN Development Program. The objective was to forge agreements on the data standards to be used to measure SDG accomplishment. The African Union Agenda 2063 was highly influential at that conference. Notably, the author of this chapter was invited to speak about Kenya's experience with KODI.

Owing to the fact that most big data are generated by non-state actors, the various national statistics offices did not consider big data to have an "official" authoritative source, noting that the notion created unnecessary distractions from their nation-centric jobs. One official said, "It will bring chaos, as there will be no control." Hence, apart from a few collaborative efforts with mobile network operators (Hackl, 2015) to provide data proofs-of-concept and scope estimations, there has been very little experimentation by African states. In May 2015, a regional OGP meeting took place in Dar el Salaam, which provided amorphous outcomes that again failed to address the delegates' fears and concerns for data proprietorship, and the official culture of the respective national offices was even more entrenched.

At the UN World Data Forum in Cape Town, South Africa, in January 2017, the author had the privilege of attending meetings on big and open data implementations, where it was found that the various related developments

made since 2015 have improved transparency and produced several productivity advancements. Moreover, the realization that open data methods can legitimately empower citizens in positive ways cast a promising light on the historical stalemate.

3.9 Why We Need Transparency and Openness

In reference to ecological systems theory (Bronfenbrenner, 1977; 1979), at the micro level, the interactions between policymakers, donor agencies, NGOs, and academics worldwide are assumed to be immediately and directly impactful to national development. Hence, the lack of impact can be ascertained as a negative social advancement. The mesosystem level involves the many interconnections among microsystems, international relations, and agency developments, which comprise powerful knowledge-sharing regimes that require transparency and openness to facilitate future capacity.

In most poverty-stricken regions of the world, basic services, such as clean water and electricity, are rare and often unreliable. Hence, the citizens of those regions remain excluded from the rest of digital society in addition to basic human dignity. Digital mapping is a concept that is changing this situation to provide data accessibility and knowledge visibility to these people. Moreover, it enables both dispossessed citizens and the larger global digital community to influence government policymakers. A compelling digital mapping case is described in the next section.

3.9.1 Mapping of Kibera, Kenya

After the city of Nairobi, Kenya, began digitizing their land maps in 2005, it was discovered that the shanty town of Kibera, with over 300K inhabitants, had not been well-represented (see Figure 3.1 for an aerial view of the area). The mappers were aware of the town's existence, but they considered the structure too complex. Hence, Kibera was excluded from infrastructural advancements, including emerging water-based e-commerce enterprises, census coverage, and many other social services. Making matters worse, some NGOs were found to have inflated the numbers (upwards of one-million) of residents to obtain excessive donor funding, resulting in a huge backlash (Karanja, 2010; Robbins, 2012) and setting back the world's eagerness to continue helping. Finally, in 2009, the national census revealed Kibera's 300K population. Nevertheless, this population persisted

Figure 3.1 Aerial view of the Kibera slum before digitization.
Courtesy of Ground Truth.

within a 2-mi^2 shantytown. Coincidentally, a new national policy limiting the number of private primary school graduate admissions to top public sector secondary schools increased the marginalization of Kibera residents, as their schools were predominantly run by charity organizations and churches (i.e., "private").

Assisted by Ground Truth, an NGO based in Washington, DC, a team of mappers from Kibera used open-source QGIS and ARCGIS tools to produce an interactive digital map of the area (see Figure 3.2). Hence, for the first time, Kibera's levels of marginalization and poverty were visible to the information consumers of digital society. Notably, only 4% of Kibera's schools were publicly run, yet they were considered "preferred" by the government (i.e., received the most funding). That is, only 4% of Kibera's youth were accounted for by the public education system.

Now that Kibera enjoys a digital footprint, it receives services that had previously been withheld due to ignorance and/or complacency. For instance, motor bikes mounted with Global Position System devices can now provide e-commerce and delivery services. Security cameras and lighted areas have also been added due to the valuable crime data acquired from digital mapping services. Moreover, education, energy, water, and ICT services are being developed. Hence, it can be said that some of the SDG goals are being addressed in Kenya. However, much more data discovery efforts of this type are needed before true progress can be measured because there is a legitimate concern that Africa's measurement of SDG goal attainment simply cannot be presently quantified.

According to de Soto (2003),

> Where it is impossible to determine who owns what, where addresses cannot be verified, where it is difficult to collect debts from individuals, where it is difficult to convert resources into money, where it is difficult to divide ownership into shares, where it is difficult to standardize and compare asset descriptions, and where the laws governing property vary from one neighborhood or even one street to another.

That is, de Soto identified several aspects of the human interactive, transitioning, and problem-solving methods identified in Bronfenbrenner's ecological theory.

In 2019, the Kenya National Bureau of Statistics established that there were 250K people living in the shantytown, which helped solidify new education policies. Moreover, very important lessons were learned about the benefits of digitization and open data in cases where hundreds of thousands of lives were

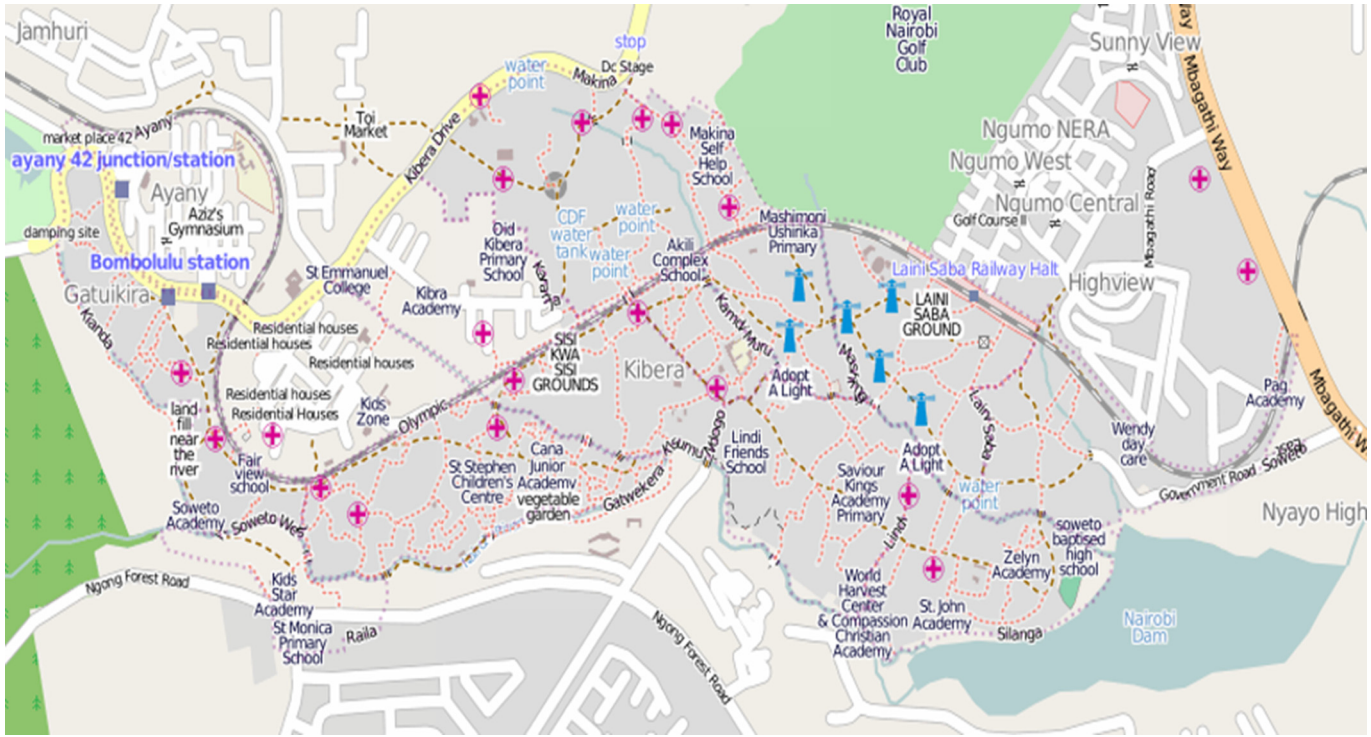


Figure 3.2 Kibera after digitization.
Courtesy of Ground Truth.

impacted positively. Hence, it is an empirical example that supplies pragmatic reasons for adopting wide-scale digitization programs (Graham et al., 2014).

3.10 Discussion and Conclusion

The impact of big and open data at the ground level must be closely attended by academia and governments that willingly apply analytics to policymaking; however, Africa's constituents remain underrepresented at global conferences, despite the author's ability to acquire sufficient information for this research. Recall that it took six years to do so. Nevertheless, it is through these types of interactions at global conferences that influential members of this continent's nations can learn from one another to gain the wisdom, competence, and confidence to leverage digital information and ICT resources to build wealth and empower citizens from within. This would also reduce the dependence on global donor agencies and NGOs, which nearly always have their own overriding visions for the continent.

Notably, the ecological framework of Bronfenbrenner (1977; 1979) is reflected in the Kibera success story. Hence, policymakers, academics, and relevant development agencies have in their hands a strong theoretical, empirical, and practical set of outcomes and measures to complete the transition to big and open data-driven SDG implementation and measurement capabilities. Very importantly, despite the Kibera example occurring in the heart of Kenya, the example demonstrates that continued collaboration with international audiences is needed. However, strong skepticism remains with the various government agencies in Africa regarding external knowledge sources, which are fueled by an ingrained need to protect oneself against larger geopolitical interests.

At the macro level, we find several authentic cultural influences that drive socioeconomic issues. Contrary to the many false perspectives owned and propagated by global audiences, the continent of Africa does not share a single cultural identity. Moreover, although there are several common and fundamental aspects in the cultural variety, more than 3,000 identifiably separate cultures exist on the continent. Europe has ~160, and North America has 11. As was clearly demonstrated in the sixteenth to twentieth centuries in Africa and in the fourteenth to ninetieth centuries in North and South America, highly fragmented cultural groups are prone to conquest. Hence, the continued fear of neo-colonialism in Africa remains legitimate.

Notably, cultural values and risk tolerance change with successive generations. Moreover, overall cultural homogenization continues to take place, and

it has been greatly accelerated with the growing ubiquity of the digital global society. The Kibera case is an exemplary historical case that is still in development. Alongside the fusion taking place in digital culture, new existential fears are being ingrained into the new generations based on data breaches, hacking, and large-scale analytical warfare (e.g., public media manipulation, election subversion, and fomentation of hate).

Therefore, as Africa inevitably invests in new data centers and a continental information infrastructure, the role of openly accessible big data remains both controversial and vital to implementing and measuring SDG progress, whose goals expire in 2030. The hesitance of African policymakers to leverage emerging ICTs for these purposes is being overcome, and the still-young twenty-first-century generations of future leaders are likely to carry the SDGs forward to fruition, if not by 2030, then by a later date that will require the continued diligence of even newer generations. With public data collection comes the existential need for public participation, meaning that the citizens of the world must become more connected to the extent that eventually, no one can possibly be left unaccounted. As such, big data analytical tools must be continually improved while making them less expensive so that they become primary social resources, such as water and electricity.

This article has demonstrated that, to carry society forward, our data analytics must be big, open, and transparent. A cogent example was found in the case of building and maintaining the Kibera digital map. Aspiring strategies are needed to mine and render the vast and unimaginably valuable veins of the big data "ore" present on the African continent. By the nature of this integration, which involves historically marginalized citizens, vast and concentrated global digital communities will be forced to become more altruistic and accountable. Hence, the shared data will hopefully be used to streamline the various statutory and regulatory frameworks that have tackled large social problems one at a time. However, individual human rights and liberties are not merely influenced by one social issue at a time. Hence, the case being made for big and open data integration in the pursuit of achieving and measuring UN SDGs is very strong. With unified awareness comes strength, and with ubiquity comes transparency. Big data comprise the world's greatest resource going forward, and Africa has vast potential wealth in this area. However, this future "wealth" is one of progressive policies that aims to improve the wellbeing, safety, and prosperity of all the world's people.

Even with this strong optimism, both the global citizenry and their various governments must remain vigilant and progressively skeptical with regard to the hopes and threats. Moreover, we should help prevent the global citizenry and/or their governments from *becoming* threats. Today, vast analytical

power is being leveraged by private citizens, firms, and governments for the purposes of perpetuating historically imperialistic and anarchist tendencies of profit and past generations. As such, surveillance is being applied in place of social welfare monitoring, and international social media platforms are being applied as propaganda mills. In summary, Africa must be powerful with its science, resolve, and leadership so that its respective nations may obtain and maintain the capacity to leverage data for prosperity while disallowing theft and/or subversion. Moreover, the application and development of the UN SDGs are designed to guide this process, which prioritizes the protection of human rights worldwide.

References

United Nations Commission for Africa (2015). Africa Data Consensus Addis Ababa, Ethiopia. www.cgdev.org/sites/default/files/Africa-Data-Consensus.pdf.

Ahmed, A. (2007). Open access towards bridging the digital divide: Policies and strategies for developing countries. *Information Technology for Development*, *13*(4), 337–361. doi: 10.1002/itdj.20067.

Bernard, H. R. (1994). *Research Methods in Anthropology: Qualitative and Quantitative Approaches* (2nd edition). Walnut Creek, CA: AltaMira Press.

Bertalanffy L. (1968). *General System Theory. Foundations, Development, Applications*. New York: Braziller

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, 32(7), 513.

Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. *American Psychologist*, 32, 513–531.

Choi, J., Hwang, M., Kim, G., Seong, J., & Ahn, J. (2015). Supporting the measurement of the United Nations' sustainable development goal 11 through the use of national urban information systems and open geospatial technologies: a case study of south Korea. *Open Geospatial Data, Software and Standards, 1*(1), 4. doi: 10.1186/s40965-016-0005-0.

de Groot, R. S., Wilson, M. A., and Boumans, R. M. J. (2002). A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecological Economics*, 41(3), 393–408. doi:10.1016/S0921-8009(02)00089-7.

De Soto, H. (2003). *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. New York: Basic Books.

Graham, M., Hogan, B., Straumann, R. K., and Medhat, A. 2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4), 746–764.

GSMA (2014) Measuring mobile penetration: Untangling "subscribers," "mobile phone owners" and "users."

Hackl, P. (2015) Big data: what can official statistics expect? conference proceedings: österreichische statistiktage 2015 at: Vienna, Austria.

Helliwell, J., Layard L., & Sachs, J. (2015). World happiness report 2015. United Nations, Sustainable Developments Goals (SDGs).

Hepworth, D. H., Rooney, R. H., Rooney, G. D., Strom-Gottfried, K., & Larsen, J. A. (2010). *Direct Social Work Practice: Theory and Skills* (8th ed.). New York: Brooks/Cole.

Huntington, H. P. (2000). Using traditional ecological knowledge in science: methods and applications. *Ecological Applications*, 10(5), 1270–1274. doi:10.1890/1051-0761(2000)010[1270:UTEKIS]2.0.CO;2. Archived from the original (PDF) on 2012-01-21. Retrieved 2011-06-27.

IEAG (2014) A world that counts: mobilizing the data revolution for sustainable development. Independent Expert Advisory Group.

Independent Expert Advisory Group (IEAG) (2014). *A World that Counts: Mobilising the Data Revolution for Sustainable Development*. New York: United Nations.

Jackson, L. A., Zhao, Y., Kolenic, A., Fitzgerald, H., Harold, R., & Von Eye, A. (2008). Race, gender, and information technology use: the new digital divide. *CyberPsychology & Behavior*, 11(4), 437–442.

Karanja, M. (2010). Daily Nation: Myth shattered: Kibera numbers fail to add up.

Kawulich B. B., (2005). Participant observation as a data collection method. *Forum: Qualitative Social Research*, 6(2), 43. https://www.researchgate.net/publication/221875465_Participant_Observation_as_a_Data_Collection_Method

Ndemo, B. and Mkalama, B. (2024). Digitalisation and Financial Data Governance in Africa: Challenges and Opportunities. In *Ndemo, Ndung'u, Odhiambo and Shimeles, Data Governance and Policy in Africa*. file:///C:/Users/BitangeNdemo/Downloads/978-3-031-24498-8%20(12).pdf.

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., and Haydock, M. (2011) MIT Sloan management review and ibm institute. the widening divide. how companies are achieving competitive advantage through analytics. MIT IBM Analytics.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52(2), 21.

Lerman, J. (2013). *Big Data and Its Exclusions*. Stanford: Stanford Law Review.

Manyika et al. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

Odhiambo, O., and Umar, F., (2019) Harnessing Big Data for Sustainable Development in Nigeria. *Journal of Sustainable Development*, 12(3). 2019 ISSN 1913-9063 E-ISSN 1913-9071. Published by Canadian Center of Science and Education.

Open Governance Partnership (OGP). https://www.opengovpartnership.org/

Robbins (2012) The Guardian: the missing millions of Kibera, https://www.theguardian.com/science/the-lay-scientist/2012/aug/01/africa-propaganda-kibera

Schmuck, R. (1997). *Practical Action Research for Change*. IRI/Skylight Training and Publishing.

Seifert A, Cotten SR, Xie B. (2021) A double burden of exclusion? Digital and social exclusion of older adults in times of COVID-19. *The Journals of Gerontology Series B Psychological Sciences & Social Sciences*, 2021 February 17;76(3):e99–e103. doi: 10.1093/geronb/gbaa098. PMID: 32672332; PMCID: PMC7454901.

Sustainable Development Solutions Network. (2015). *Data for Development: A Needs Assessment for SDG Monitoring and Statistical Capacity Development*. Sustainable Development Solutions Network.

UNECA. (2022). Africa SDGs Progress Dashboard. https://sdg.iisd.org/news/uneca-launches-africa-sdgs-progress-dashboard/

4

Making Human Measurement Matter for Human Development

Towards a Theory of Change for the Data Revolution and "Human AI"

Emmanuel Letouzé

4.1 Introduction

Does accurate measurement matter for the achievement of outcomes being measured, the way most scholars believe that institutions matter, i.e., have a causal effect on observable outcomes (Acemoglu & Robinson, 2012; Przeworski, 2004)? How could measurement matter more in the age of data and AI?

Whether termed "the Industrial revolution of Data" (Hellerstein, 2008), "Computational Social Science" (Lazer et al., 2009), "Big Data," the "Data Deluge" (The Economist, 2010), or simply the "Data Revolution" (United Nations, 2013), an era of rapid socio-technological change started at the end of the first decade of this century that would, according to some, no less than "transform how we live, work and think" (Mayer-Schönberger & Cukier, 2013), raising both hopes and fears. A decade later, "Artificial Intelligence" (AI) has replaced "Big Data" in the public discourse and is also—all the more since the release of Large Language Models (LLMs) such as ChatGPT—expected to change just about everything. Despite differences between these concepts, they are all parts of the same phenomenon, a Data Revolution (or Data *and* AI Revolution), stemming from the rise of digital data and tools in our lives that gives us an unprecedented ability to quantify many of their aspects and facets.

In the so-called international development community, most of the discussions and excitement surrounding the Data Revolution soon focused on ways in which these new kinds of data and analytical methods might fill "data gaps" and, in particular, help better measure, monitor, and promote, the

Emmanuel Letouzé, *Making Human Measurement Matter for Human Development*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press.
© Emmanuel Letouzé (2025). DOI: 10.1093/oso/9780198879145.003.0005

17 Sustainable Development Goals (SDGs), their 169 targets, and 232 indicators, agreed upon by 193 Heads of governments in 2015 (United Nations, 2017).

Whether made implicitly or explicitly, a core argument is that the dearth of quality data and in particular of good official (or public) statistics—sometimes referred to as a "statistical tragedy" (Devarajan, 2011)—has played a major role in holding back progress. In other words, the argument is that the fact many human development outcomes such as poverty, literacy, mortality, health etc., including those captured by SDG indicators are badly or hardly measured explains to a significant degree the bad state of these outcomes. According to this logic, alleviating this constraint would lead to measurable improvements in the outcomes being assessed. One of the clearest expressions of this line of argumentation is the phrase "better data, better decisions, better lives" and its variations,[1] which *do* suggest that some *causal* process does (or should?) take place.

There is no denying that better human development statistics can help promote human development; that better measurement (of income poverty for example) can help affect what is measured (i.e., lower poverty). However, the strength, steps and enabling conditions of this causal link are hardly ever spelled out, much less scrutinized. Lacking in discussions about the power of data to create change are both a critical assessment of the causal claim on the basis of available evidence, and a clear theory of change of how these measurements may causally foster human development; in other words: how can measurement *matter*, or matter more, in a causal sense, in the age of data and AI? Such is the main objective of this chapter.

I start by presenting the general context, core concepts, and standard arguments around the Data Revolution narrative of the past decade. I then discuss the validity and limitations of that view, focusing on the measurement–development link through historical and empirical perspectives. Finally, I focus on a vision termed "Human AI" and propose to expand on it with a theoretical and a practical contribution. For the former, I develop a theoretical–mental framework that offers a theory of change linking development data and AI to development outcomes and for the latter, I propose three functions of data, which will give practical and actionable pointers to those seeking to adopt data and AI into their work. Together these contributions theorize and operationalize the vision of Human AI and may enhance the causal relationship between measurement and development—i.e., *make data and measurement matter (more) for human development* by enabling societal learning and improvement.

[1] See for example (OHCHR, 2014; Bullinger & Barnes, 2015; GPSDD, 2018; Data2X, 2023).

4.2 The Pillars and Promise of the "Data Revolution": "Better Data, Better Decisions, Better Lives"

4.2.1 The Dawn and Hopes of a New Era

Among the first to ring the bell heralding the arrival of the Data Revolution a decade ago were the authors of a UN report on the Post-2015 Development Agenda that called for a "data revolution for sustainable development" (United Nations Secretary-General, 2013) (see Box 4.1), followed two years later by another titled "A World That Counts: Mobilizing the Data Revolution for Sustainable Development," which opened with the line "Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time, effective policies become almost impossible." Dozens of papers, books, and conferences on this topic soon followed, thoroughly embedding this notion into popular discourse within the development community and beyond. Notable among them was an ODI report from 2015, "The Data Revolution: Finding the Missing Millions" (Stuart et al., 2015) and UN Global Pulse's White Paper (which I authored) from 2012 "Big Data for Development: Opportunities and Challenges" (Letouzé, 2012).

This excitement came on the heels of the United Nations' Millennium Development Goals (MDGs), in place from 2000 to 2015 and grew with the new Sustainable Development Goals (SDGs), active from 2015 to 2030 (UNDP, 2023a). By design, both the MDGs and SDGs reflect the notion that measurement fosters development, echoing the common view that "what you can't measure, you can't manage." More precisely, the SDGs are meant to provide an overall vision (the SDGs Agenda) and yardsticks for measuring progress (the SDG targets and indicators) towards it.

Under what I call the standard Data Revolution logic and line of argumentation, policymakers and world leaders with poor or no data are like airplane pilots flying blind without instruments and metrics (Giugale, 2013). Improving the quality of data (making them more granular, reliable, timely, disaggregated, and relevant) will likely ensure that economies and societies reach their destinations. And the Data Revolution could help.

The advent of "Big Data" in particular seemed to provide new, previously unthinkable, opportunities to "fix" or "fill" the gaps in official statistics and SDG monitoring, initially framed as complements to official statistics and increasingly as inputs to and integral parts of the statistical production process. Early research showed that basic development statistics such as income poverty could be estimated from new data sources such as cell phone metadata (Blumenstock et al., 2015) or satellite imagery (Noor et al., 2008),

as well as several other SDG Tier 1 indicators (see Box 4.1). The possibility of drawing on these and other new types of data sources such as social media and transaction data to help measure and monitor indicators that are less routinely available, known now as SDG Tier 2 indicators, was soon another appealing promise (Data-Pop Alliance, 2019). (See Box 4.2.)

Box 4.1 A New Data Revolution

"Too often, development efforts have been hampered by a lack of the most basic data about the social and economic circumstances in which people live . . . Stronger monitoring and evaluation at all levels, and in all processes of development (from planning to implementation) will help guide decision making, update priorities, and ensure accountability. This will require substantial investments in building capacity in advance of 2015. A regularly updated registry of commitments is one idea to ensure accountability and monitor delivery gaps. We must also take advantage of new technologies and access to open data for all people."

Source: Bali Communiqué of the High-Level Panel, March 28, 2013

Box 4.2 Tier Classification Criteria/Definitions

Tier 1: Indicator is conceptually clear, has an internationally established methodology and standards are available, and data are regularly produced by countries for at least 50% of countries and of the population in every region where the indicator is relevant, such as indicator 1.1.1. measuring extreme income poverty.[a]

Tier 2: Indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries, such as indicator 1.4.2. monitoring land rights.[b]

As of November 30, 2022: The updated tier classification contains 148 Tier 1 indicators, 77 Tier 2 indicators and 6 indicators that have multiple tiers (different components of the indicator are classified into different tiers).

[a] Indicator 1.1.1: "Proportion of the population living below the international poverty line by sex, age, employment status and geographic location (urban/rural)."
[b] Indicator 1.4.2: "Proportion of total adult population with secure tenure rights to land, (a) with legally recognized documentation, and (b) who perceive their rights to land as secure, by sex and type of tenure."

Source: (United Nations, 2022a). IAEG-SDGs: The Classification for Global SDG Indicators. Sustainable Development Goals. https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/

4.2.2 Promises and Lessons of Big Data for the Data Revolution

This standard Data Revolution narrative did and does recognize that better or bigger data won't mechanistically and seamlessly fix the world's most deeply entrenched problems. As early as 2013 (see Box 4.1), the need for strengthening capacities and accountability mechanisms were mentioned. Major ethical questions related to privacy and power were also quickly identified in the early years, with a focus on those raised by Big Data (Boyd & Crawford, 2011; King, 2011; Letouzé, 2012).

Whereas the Data Revolution cannot be restricted to Big Data, Big Data fueled much of its initial hopes and fears, and debates about the nature, novelty, and challenges of Big Data help us think through the assumptions, risks, requirements, and limitations of the Data Revolution more broadly. Big Data still does not have a commonly accepted definition, but it is most often used as an "umbrella term" (Letouzé, 2012) that refers to both the large-scale data (i.e., big data) on social outcomes created and collected through digital devices and services, and the computational capacities and practice of analyzing them. As Gary King put simply, "Big data is not about the data, it's about the analytics" (2016). Former Amazon Chief Data Scientist Andreas Weigend meanwhile defined Big Data as "a mindset, to turn mess into meaning" (TSIA, 2013).

Starting from 2013, I conceptualized Big Data as a complex ecosystem distinct from big data (and their "3Vs," volume, velocity, variety, or more Vs)[2] serving at its core inputs or fuel, now made up of "4 Cs," which I proposed to consist of Crumbs, Capacities, Community, and Culture, depicted through four concentric circles, each covering a piece of the phenomenon (see Figure 4.1). While the framework was developed for Big Data, I propose that it can be used to capture and analyze the constitutive elements of the Data Revolution more broadly.

Importantly, there are shades of gray and shifts in what data qualify as Crumbs. Initially, I equated Crumbs with "big data" such as call detail records and social media data, but this can be debated. First, census microdata, administrative data and data collected by humanitarian agencies on recipients are large sensitive datasets and, as such, they may be considered as big data. Further, as new kinds of data get created and collected or transformed digitally, what is deemed big data will evolve. While the core, the fuel, of Big Data and AI are those digital records passively emitted about every

[2] For a discussion on the Vs see Letouzé, Meier and Vinck (2013).

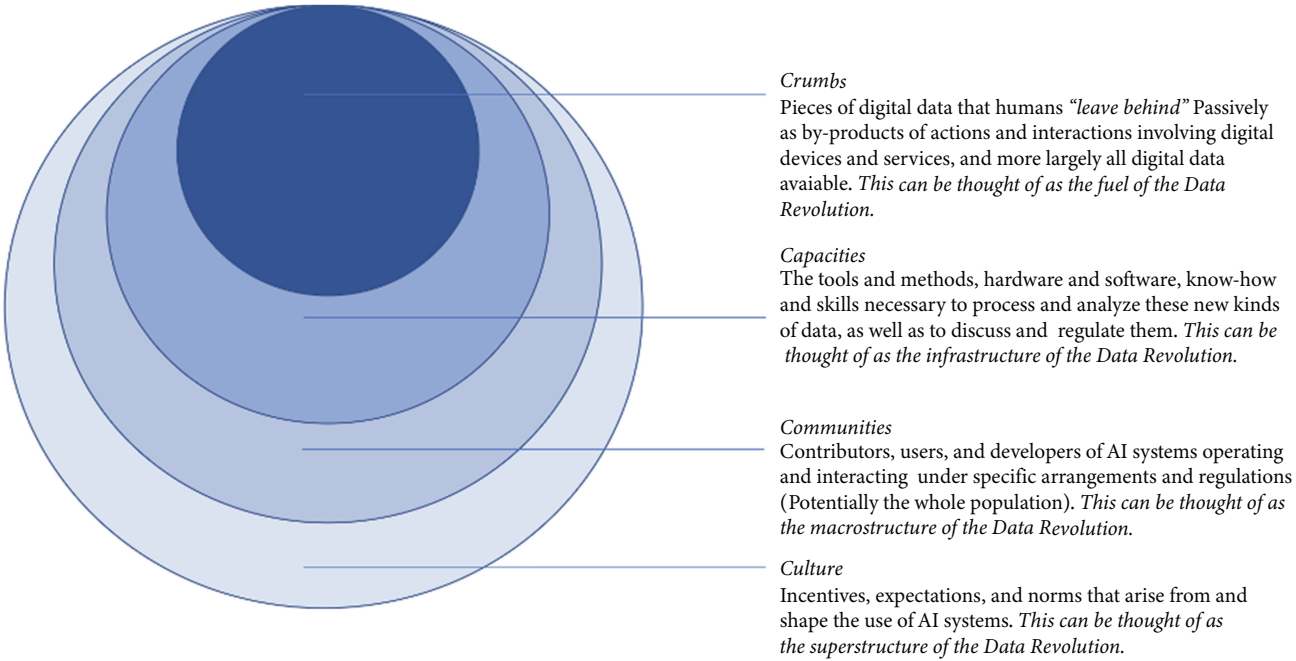


Figure 4.1 The 4 Cs of Big Data, AI, and of the Data Revolution.

Reproduced with permission from Letouzé, E., Oliver, N., Lepri, B., & Vinck, P. (2023). AI for the SDGs: And beyond? Towards a human AI culture for development and democracy. In B. Prud'homme, C. Régis, & G. Farnadi (Eds.), *Missing Links in AI Governance*. United Nations Educational, Scientific and Cultural Organization (UNESCO).

one of our actions and interactions, it is likely that in some non-distant future the core will be simply considered as "data" of different types and attributes—including official statistics as a type of data produced according to Principles of Official Statistics, including using non-traditional data as inputs thanks to the appropriate Capacities.[3] Today, I consider Crumbs to encompass all types of digital or digitizable data, i.e., nearly all existing data points.

4.2.3 15 years of a Data Revolution: A Mixed Record at Best

Taking serious stock of some 15 years of Data Revolution would take much more time and space than this chapter allows, but at least two general observations can be made.

First, the promises of the Data Revolution have not, to date, been fully fulfilled; its record is mixed at best. There have been progress and successes—including in using data to respond to the COVID-19 pandemic. The pandemic did not help, but it is hard to argue that the world of 2025, with civil and international conflicts, accelerating climate change and environmental degradation, and ever-growing inequalities, is markedly better than that of the pre-Data Revolution era; say before the 2008–09 financial crisis, despite the ever-increasing amounts of data. In fact, recent years have seen backslides on measurement capacity, slowdowns or reversal of progress towards core SDGs such as mortality rates in several countries, and democratic decline worldwide since 2007 (Hille, 2022). Some even argue that "data killed facts" (Lepore, 2020), suggesting that the current state and trajectory of the Data Revolution since the invention of the Web may even run counter to its initial promise of democratization. At minimum, we should recognize that we could have expected and still ask more, from the Data Revolution.

Second, indeed, many questions remain about the potential of the Data Revolution to contribute to sustainable human development by improving measurements. A key problem that the 4C framework highlights is the hiatus between the standard, rather mechanistic, Data Revolution line of argumentation, and the actual functioning and complexities of the world: whereas many have discussed the importance of data (Crumbs) and of the abilities to analyze them (Capacities), little mention has been made about the importance of communities and culture, as defined here, i.e., issues of power, politics, trust, values, and well as psychology. The full potential of the

[3] For fuller discussions see Letouzé (2012) and (2015), as well as Letouzé, Oliver, Lepri, and Vinck (2023).

Data Revolution is hampered without a culture that sees value in data and AI or communities that advocate for and adopt data and AI, while remaining vigilant for and seeking to prevent their potential harms.

That facts have a hard time changing our minds and actions on an individual level is well known (Kolbert, 2017). For example, a speed loving driver of a fast car will be unmoved by the unambiguous signal provided by his speedometer that he is indeed speeding.[4] So is the fact that politics and economics often do trump statistics when it comes to collective decisions, or decisions with collective implications. But the implications of these facts have insufficiently translated into a fuller, deeper, theory of change and action plan for the Data Revolution. Before attempting to propose one, I want to delve a bit deeper into the complex relationship between measurement and development, from a two-pronged perspective coupling historical and empirical viewpoints.

4.2.4 Revisiting the Measurement-Development Nexus in a Historical and Empirical Perspective

4.2.4.1 An Abridged Critical Historical Analysis of Measurement and Development

Over most of the past 300 years at least, measurement and human development have been intertwined in complex ways (Vincent, 2022). Before the Industrial Revolution and the subsequent demographic transition, about all we know is that, as put by Ronald D. Lee (2003), "life was short, births were many, growth was slow and the population was young." While key demographic indicators such as births and deaths were recorded, usually by religious institutions such as churches in Europe, most modern economic indicators had not been invented yet. This slowly changed when statistics, from the german *Statistiks*, or "Science of the State,"' took a turn as emerging European nation states started systematically compiling records of outputs and stocks of goods, in large part for warfare and tax collection purposes (Stigler, 1986).

During this time, scientific and technological progress that gradually infused daily lives led to increases in living standards over time—as captured for example through life expectancy at birth, which was, in retrospect, estimated to be 28.5 years globally in 1800 (Riley, 2005). Improved data collection, hypothesis testing, and so forth, in many fields, underpinned much

[4] I make the assumption that most speeding drivers are males.

of the initial waves of the Industrial Revolution, although at the time it led to initial reversals in development outcomes such as literacy among the working class (West, 1978). Yet, it was not until the 1930s, well over 150 years after economic growth had begun to accelerate markedly, that its key measure—the Gross Domestic Product (GDP)—was invented (Dickinson, 2011). In any event, the ability to quantify and study processes scientifically on the basis of measurement seems to have played a major role in recent human progress.

GDP is a great case to study the nature and limitations of the link between measurement and outcomes (understood as changes in what is measured). It was created as part of the development of the System of National Accounts (SNA) in the wake of the Great Depression to assess whether macroeconomic policies implemented in the 1930s were effective. Since then, GDP has become and to this day remains the alpha and omega of policymaking and makers; used interchangeably with economic activity. Over the past 90 years, whether or not a given economic policy was deemed to have "worked" or not has been assessed first and foremost by researchers and politicians alike in light of its effect on GDP, or just on "growth."

An interesting insight is that for a long time measuring economic activity (through GDP) was largely unnecessary for its promotion, although given the absence of a counterfactual it is impossible to tell whether living standards would have grown faster sooner had they been measured better earlier. But does this mean GDP as a metric has not mattered? Not quite.

GDP has had its critics early on, especially as it became perceived as something it was not intended to capture: welfare. In the 1970s and 1980s, growing concerns for the shortcomings and negative effects of mass consumption led to various proposals, ranging from degrowth to alternative measures of economic activity (Georgescu-Roegen, 1971, 2020). The Human Development Index (HDI), which accounts for income per capita (GDP) as well as educational and health (life expectancy) achievements, has been one of the most successful of such measures (UNDP, 2023b). Although the latter—health and education—tend to be highly correlated with the former, their inclusion has served important signaling and advocacy purposes. Measuring and including health and educational outcomes meant that these had intrinsic value, i.e., that they mattered, and should be promoted notably through the instrumental role of their measurement.

The role of measurement is also suggested by what happens to outcomes that *cannot* or *are not* measured, which are by definition or design statistically invisible. As suggested above, the ecologically detrimental effects of pro (GDP) growth policies went largely ignored for decades. Once they became a significant concern, through anecdotal evidence and other measurements,

new sustainability indicators were invented, some of which capture fiscal sustainability that take into account national savings and investment. Other measures such as "Adjusted Net Savings" were created and used to take into account other processes and outcomes missed by standard macroeconomic indicators (Gnegne, 2009). All of this makes it safe to assume that, to a large extent, GDP is and has been one measure that has "mattered"—for better and worse, the same way it is argued that institutions *matter* (Przeworski, 2004).

It was in this context that the HDI then the MDGs were established to add yet more indicators that should be taken into consideration and measured. The MDG agenda had mixed results, both in terms of measurements and achievements, although it is near impossible to estimate how much of the latter can be attributed to the former (McArthur & Rasmussen, 2017). The subsequent SDG Agenda, considered to have been devised in a more participatory manner, expanded its reach to new topics. The SDGs reflected the core tenet of the Data Revolution and of public statistics more largely that societies measure what matters to them, and that this very action, in turn, ought to matter.

A key simple message is that statistical measurements are *socio-political statements*—not natural objects, but statistical constructs that mean "this outcome matters, hence it should be measured." But how and how much measurements actually matter remains a big question.

4.2.4.2 A Quick Empirical Analysis

In an attempt to further investigate the link between measurement and development (outcomes), we can analyze two main statistical estimations of their levels: for the former, we take the World Bank's Statistical Capacity Index (SCI) (2023), computed on a 0–100 scale based on 25 indicators. For the latter, we take UNDP's HDI (2023b), which takes into account life expectancy, education, and income on a 0–1 scale. The hypothesis is that if statistical capacity was a good proxy for statistical quality, which I think it is, *and* if statistical quality mattered very much for affecting the outcomes measured, then we should find a strong correlation between SCI and HDI.

Running the analysis for all 140 countries for which the SCI is available for 2015, the starting point of the SDG Agenda, shows a statistical relationship between SCI and HDI analyzed (average P-value ~ 10^{-7}) and yields an R^2 of about 0.19 (see Figure 4.2). This means that differences in SCI account for less than 20% of the variation in HDI. The slope estimate suggests that a 10-point change in SCI (ranging from 0 to 100) is associated with a 0.03 point change in terms of HDI (ranging from 0 to 1) (see Table 4.1). We do see a generally positive relationship, but also lots of variations in the sample.

In short, statistical capacity (here a proxy for measurement) does seem to matter, but not very much.

A few specific observations or hypotheses can be developed on the basis of this simple analysis about the strength, nature, as well as the historical, and political facets of the relationship between SCI and HDI.

First, most countries with high and very high HDIs (above 0.7) also have high SCIs (above 80). However, there are outliers with low SCIs and high HDIs such as the Small Island Developing States (SIDS) of Saint Kitts and

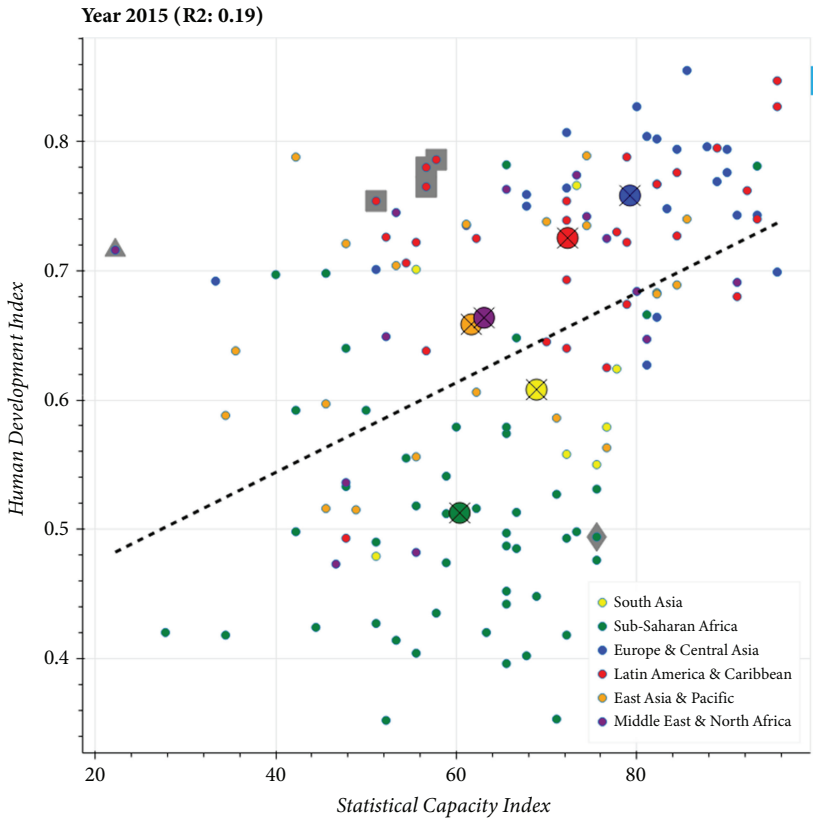


Figure 4.2 Scatter plot of SCI vs HDI for 140 countries by region, year 2015.

Table 4.1 Regression results for SCI and HDI on 2015.

	Coefficient	Std error	P>|t|
Intercept	4.05×10^{-1}	4.32×10^{-2}	8.61×10^{-17}
Slope	3.47×10^{-3}	1.61×10^{-3}	1.13×10^{-7}

Nevis, Antigua and Barbuda, Trinidad and Tobago, Grenada and Palau (highlighted with gray squares). This pattern may have roots in some common features of SIDS, including a large "brain drain" and limited economies of scale in building and running public administrations (de la Croix et al., 2014), that have led to calls for more investments in their statistical systems (United Nations, 2022b). Other countries with large natural resources displayed high HDIs and low SCIs that year, such as Libya, with its small population, vast oil resources, over 40 years of dictatorship under the Gaddafi regime, and the civil conflict and institutional crumbling that followed its fall, highlighted in a gray triangle.[5]

Other interesting cases are countries with low HDI and high SCI, including several sub-Saharan African countries. One example is Senegal (highlighted in a gray diamond), with an HDI of 0.4 and an SCI of 75 in 2015. Senegal became the fourth Sub-Saharan African country to adhere to the IMF's Special Standard for Data Dissemination (Norme Spéciale de Diffusion des Données (NSDD) in 2017, and has had an Independent Statistics Agency (Agence Nationale de la Statistique et de la Démographie, ANSD) since 2005, with its own training school (ENSAE Senegal) and enjoyed 60 years of relative peace and democracy, strong institutions, qualified human resources, a vibrant press, partnerships with and support from international institutions such as UNFPA, the French INSEE, etc. Former Soviet Union countries tend to have higher SCI than their HDI would suggest.

Of course, some readers may think "correlation does not imply causation." But the point is that the standard Data Revolution narrative "better data, better decisions, better lives" *does* suggest, or calls for, some causation, and causation is usually manifested in correlations. There are obviously lots of confounding factors mentioned above—chief among which are history, politics and culture. It can be argued for instance that some countries are somewhat culturally wired to have both good development statistics and good development outcomes. It is also evident that there is reverse causality: high income countries can afford better statistics; they also tend to have free press and civil societies that demand accountability through reliable statistics.

But the main point is simple: overall, measurement doesn't always matter very much, for various reasons mentioned above: politics, psychology, inertia, economic interests, power dynamics, and so forth. Most fundamentally, if one believes in using measurement (as a form of evidence) to inform collective decisions and policies to foster progress, then measurement should

[5] While the country's HDI started decreasing in 2012, it remained then in the very high HDI group by 2015; over the same period its SCI further deteriorated from a low 28 in 2008 to 22 in 2015.

matter more. Another way to look at it is: assume an exogenous shift to the right of a given country along the SCI axis. How can this translate faster if at all into a move upward along the HDI axis? How can that work?

4.3 Towards a New "Human AI" Theory of Change to Make Measurement Matter: Learning Loops, Lens, Language, and Lever

I go back to the core question: how, in the age of data and AI, can measurement of social phenomena matter more? How can more rigorous measurement causally address major social problems such as those covered by the SDGs among others?

I aim to contribute to partially answering this core question of the Data Revolution by extending the vision of "Human AI" proposed in previous publications—where, in summary, AI is considered and used both as an *instrument* to provide better measurements (through AI) and as an *inspiration* to use these measurements as feedbacks enabling social learning (in reference to machine learning)—with a theoretical and a practical contribution (Letouzé et al., 2023; Letouzé & Pentland, 2018).

I theorize the Human AI concept further by proposing a three-part "theory of change" centering on what I term "learning loops." I then couple this with a more practical contribution outlining "Three functions of data," connect to the aforementioned 4Cs of data and AI, which I think can help those working with development to act on and operationalize the ideas that the Human AI vision encompasses.

4.3.1 Learning Loops: Artificial Intelligence as an Instrument and an Inspiration for Learning Societies

As the name "learning loops" implies, drawing from my extensive work with Data-Pop Alliance and collaborations with a range of stakeholders in the data and development space as well as the literature on cognitive biases, I contend that human progress does not happen linearly but through loops. A learning process could be conceptualized similar to the curve depicted in Figure 4.3, which can be traced and thought of as a function of performance (P) and confidence (C). This model is inspired by the cognitive bias known as the "Dunning-Kruger effect" depicted in Figure 4.3 but differs in two main ways.

First, I place Performance (instead of Competence) on the Y axis, and Confidence on the X axis. Secondly, and more importantly, in my loop model, by

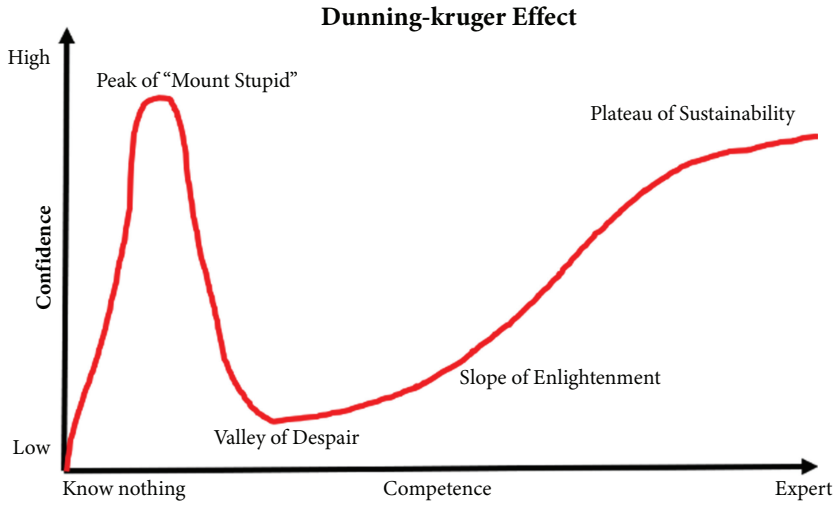


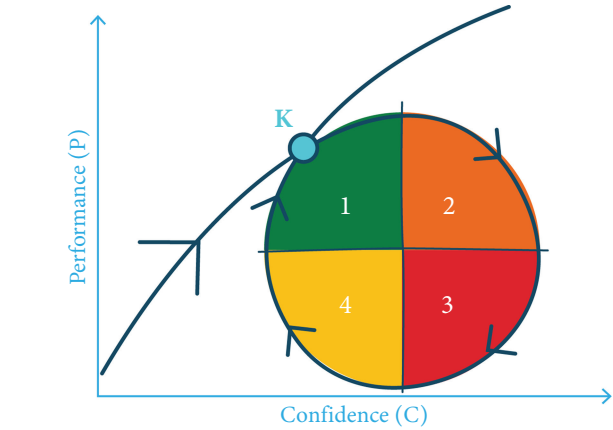
Figure 4.3 An illustration of the cognitive bias, termed the Dunning-Kruger Effect.

Courtesy of Conversion Uplift. (2023). Dunning-Kruger Effect. *Conversion Uplift*. https://conversion-uplift.co.uk/glossary-of-conversion-marketing/dunning-kruger-effect/

definition, P can decrease.[6] This latter feature allows an accounting for setbacks, such as declines in life expectancy at birth for example. I think this mental model can also account for individual activities such as learning to walk, which I will use to illustrate the process.

Next, I propose that these are feedback loops, along which learning happens that helps the individual or group of interest get back on an upward trajectory towards a higher level of P through four quadrants and steps (see Figure 4.4). The first of these is called the "consolidation phase" wherein P and C are positively correlated. For example, the child has begun to walk with ease and is gaining dexterity and strength. Before long, however, comes the second phase: overconfidence. Here, an excess of confidence has led the learner to overestimate their potential. The child may try to trot and find themself tripping over their feet, unaccustomed to this new pace. The third phase is "crisis," wherein the learner is dispirited from their recent perceived failure and loses confidence as well as ability. Afraid of falling, the child may revert to crawling which feels safer. Finally, over time, the learner enters the fourth phase, "redemption," and begins to improve in performance despite continually decreasing faith in their abilities. Just when it seems all hope is lost, the learner begins to realize the improvements made over the last phase and regains confidence, reaching the apex of the learning loop at point K, on

[6] Mathematically thus is not a function since one value of X can be associated with 2 values of Y.



1. Consolidation Mode:
P and **C** are positively correlated.
At the beginning , both performance and confidence go up.

2. Overconfidence Mode:
P and **C** are negatively correlated.
Too much confidence leads to a decrease in performance.

4.Redemption Mode:
P and **C** are negatively correlated.
Slowly, performance improves again.

3. Crisis Mode:
P and **C** are positively correlated.
During crises, both go down.

5. Inflection point K (for Kink):
Positive momentum allows the system to keep going up.

Figure 4.4 The "Learning Loops" theory of change model.

an upward trajectory. Between each learning loop is a period of relative stasis where both variables improve steadily until a novel challenge or possibility arises, plunging the system once again into a cycle of challenge. This particular example may not be the best but in my view, all individuals (including athletes and artists) and human groups go through similar processes.

This model is consistent with two observations: one, the fact that human progress in many domains is logarithmic in the long term—meaning that early progress is fast, sometimes exponential initially, while the speed of growth tapers off after some point, where returns become harder and smaller, as in the cases of GDP growth and life expectancy for example [7]—and two that human progress (P) may not be bounded.

[7] For further details, please see: Roser et al. (2013) and and Papaeorgiou & Johnson (2019). This is an important point that will warrant further work. The notion of logarithmic progress is core to the idea of economic convergence, whereby initially poorer countries may tend to grow faster (e.g., India, China, Vietnam) than richer countries (e.g., OECD members) and this "catch up" with them. Similar observations can be made in sporting events records, such as running, swimming, etc.

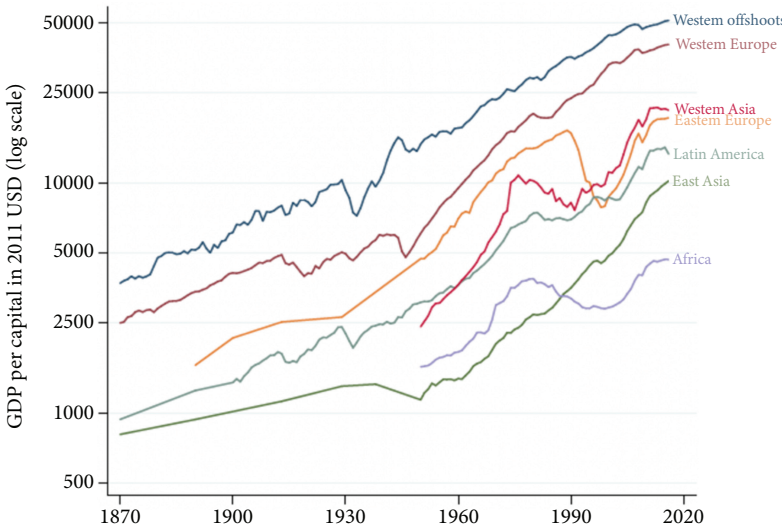


Figure 4.5 An example of generally logarithmic progress: average real GDP per capita across regions, 1870–2016.

Reproduced with permission from Bolt, J., Inklaar, R., de Jong, H., & Luiten van Zanden, J. (2018). *Maddison Project Database 2018, Rebasing "Maddison": New income comparisons and the shape of long-run economic development* [dataset]. Courtesy of VoxEU.

One key assumption of the model is actually that for progress to take place, human systems *must* go through periods of regressions (or recessions, as in Figure 4.5), whose duration varies—and could be very small and hard to see over long time series—where learning happens.

Next, as the description of the quadrants suggest, I think that data play a key role in enabling these learning loops. More precisely, I think that there are three main functions of data that are at play and that allow such a trajectory to happen: as lens, as language, and as lever.

4.3.2 Three Functions of Data for Making Measurement Matter

Through the research and other activities carried out at Data-Pop Alliance, I have come to work on data and AI with a range of stakeholders hailing from the public, private and civil society spheres. It is reflecting on these experiences that I propose the concept of "Three Functions of Data." I call these functions "data as lens," "data as language," and "data as lever"

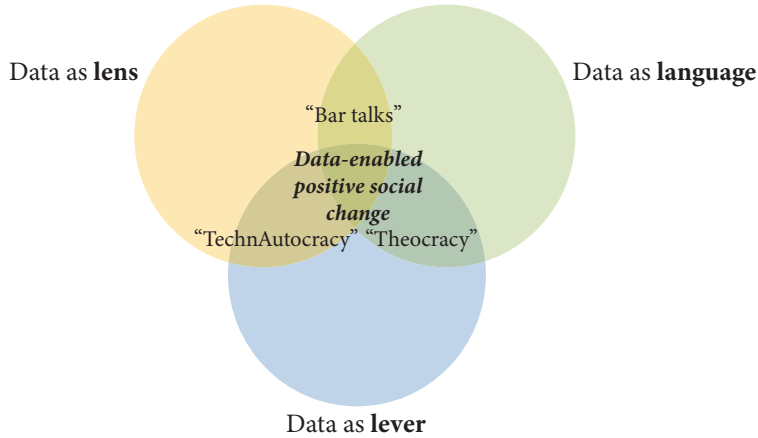


Figure 4.6 The three functions of data.

(see Figure 4.6).[8] In my view, if data serves these three functions properly, then human systems are better able to journey through these loops, and measurement can matter more.

In this section, I outline what the three functions of data entail and imply—and how they relate to the 4Cs of data and AI previously introduced. I believe that complementing the more theoretical angle of the learning loops with a practical and operational view that the three functions of data delineate, may help development actors to design actionable steps towards adopting data and AI in their work and ultimately reach better development outcomes.

Data as *lens* means using all available data (safely, ethically) to provide as good and unbiased pictures of human and natural outcomes as possible, in timely, disaggregated manner; this requires notably having access to sensitive private data in privacy enhancing and participatory ways (Letouzé & Oliver, 2019). Here, the concepts of Crumbs and Capacities, but also aspects of Communities and Culture, are key. The use of the words "picture" (of human reality) and of "lens" are telling; picture or photography is not neutral; it reflects both what the photographer chose to capture, the way societies choose what to quantify, and some subjective perspective about it.[9]

Data as *language*, has also roots in the 4Cs, especially *Culture, but also Communities and Capacities*. Using data as language refers to a context where individuals and groups are willing and able to discuss about and through

[8] Also of interest, see Jake Porway's (2021) discussion of three uses of data as "Observe," "Reason," and "Act".

[9] For an interesting discussion on the subjective nature of photography, see: (Mercier & Aubert, 2023).

measurement. In this scenario, people are not only aware of information gained through data, interested in discussing and reasoning on their basis, and demanding that things that matter to them be measured. Take, for example, the use of corporal punishment on children. Decades of studies have shown this practice not only fails to improve children's behavior, but actually damages the child's development (Cuartas et al., 2021). These research outputs would mean nothing if not for a society willing and able to use them to have reasoned discussions to evaluate and, hopefully, abolish what was once a common practice among caregivers. For data to function well as a language, there ought to be a decent level of respect between groups and trust towards institutions producing public statistics to nurture educated productive discussions. This echoes the conceptualization of democracy as "government by discussion."[10] It requires building broad-based statistical and data literacy, rekindle trust between groups, including through greater social mixing, which will require social innovations to enable it.

Finally, data can act as a transformative *lever* to enable citizens, communities, and the planet to make and implement collective decisions with measurable targets. Acting as checks and balances for the system, this component ensures that learners answer to a form of accountability for their actions. While any reputable development practitioner would perform monitoring and evaluation (M&E) of their projects, often missing are the complementary components of accountability (A) up to impact, and learning perspectives (L) to generate meaningful, up-to-date knowledge to support better decision-making. All together these four components make a more robust responsibility framework known as MEAL (Data-Pop Alliance, 2022). By ensuring decision-makers are accountable for the outcomes of their policies and learn from the process, they as learners are able to break out of the learning loop and be propelled into the next cycle. Likewise, this will require social innovations to favor accountability mechanisms to be more effective. Overall, a key ingredient for data to work as a lever is nurturing and sustaining a culture and communities where people are willing and able to use measurement to favor their own goals, in ways that enable learning loops to take place.

These three functions of data need not happen in any given order; nor do they need to take place simultaneously. But I think any situation where data is used in one or two ways and not all three at some point in some way will always yield sub-optimal outcomes.

[10] As conceptualized in different ways by J.S. Mill, John Dewey, and Amartya Sen.

4.3.3 Technological Implications and Social Innovations for Making Measurement Matter

Turning this vision into reality, or even making slow progress towards its realization, will take significant efforts in the face of multiple obstacles, including the incentives and resources of "bad actors." But a better use of data and AI is possible. It can and must be supported by targeted investments and innovations that favor inclusion and statistical literacy; by the systematization and scaling of privacy-enhancing and participatory mechanisms to access and analyze private sector data in safe sustainable ways (Eurostat, 2022), by advocacy and education campaigns about the value of information and the risks of dis/misinformation. Fundamentally, it is about convincing people that armed with data and AI, they can change the world for the better.

This needs to be quickly backed up and implemented through concrete initiatives seeking to reflect and implement the Human AI vision and functions of data described in this chapter to make measurement matter more for human development. One such initiative is the OPAL (Open Algorithms) approach, co-founded and deployed by Data-Pop Alliance (2018), built around privacy-enhancing technology and a participatory governance mechanism around the CODE concept (Letouzé & Yáñez Soria, 2021) to use sensitive data to measure and promote the SDGs and official statistics. Another is the World Bank-funded "Mobile Data for Official Statistics" Funding Facility. But a key message of this chapter is that significantly more conceptual and concrete work needs to be done for data and AI to play the transformative role that the Data Revolution advocates claim they can play.

4.4 Conclusion

This chapter has investigated the complex and evolving relationship between measurement and human development. While many consider the Data Revolution to linearly lead to better development outcomes by virtue of enabling better planning, monitoring, and evaluation of development interventions and projects, I have elucidated that many of these promises remain, in fact, unfulfilled. With a goal of better understanding this dynamic, I have proposed a theoretical and practical contribution to the vision of Human AI. The Learning Loops theory of change that I propose deepens the theoretical grounding of the concept of Human AI while the Three Functions of Data that I outline allow those working in development to consider practical and actionable ways to implement the Human AI vision in their work. Offering these insights will support development actors interested in integrating

data and AI to strengthen the activities related to measurement in their work, and it is my hope that this may ultimately contribute to achieving improved development outcomes.

References

Acemoglu, D., & Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Profile Books.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Bolt, J., Inklaar, R., de Jong, H., & Luiten van Zanden, J. (2018). *Maddison Project Database 2018, Rebasing "Maddison": New income comparisons and the shape of long-run economic development* [dataset]. https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2018

Boyd, D., & Crawford, K. (2011). *Six Provocations for Big Data* (SSRN Scholarly Paper 1926431). https://doi.org/10.2139/ssrn.1926431

Bullinger, J., & Barnes, R. (2015, March 12). *Better data, better lives: Chairing the UN Statistical Commission—Civil Service*. https://civilservice.blog.gov.uk/2015/03/12/better-data-better-lives-chairing-the-un-statistical-commission/

Conversion Uplift. (2023). Dunning-Kruger Effect. *Conversion Uplift*. https://conversion-uplift.co.uk/glossary-of-conversion-marketing/dunning-kruger-effect/

Cuartas, J., Weissman, D. G., Sheridan, M. A., Lengua, L., & McLaughlin, K. A. (2021). Corporal punishment and elevated neural response to threat in children. *Child Development*, *92*(3), 821–832. https://doi.org/10.1111/cdev.13565

Data2X. (2023). *Better Data, Better Decisions*. Data2X. https://data2x.org/resource-center/better-data-better-decisions/

Data-Pop Alliance. (2018, August 21). *OPAL for Public Data and Good*. Data-Pop Alliance. https://datapopalliance.org/transform/opal/

Data-Pop Alliance. (2019, March 27). *Measuring the Unmeasured: Tier 3 SDGs*. Data-Pop Alliance. https://datapopalliance.org/transform/measuring-the-unmeasured/

Data-Pop Alliance. (2022). *WIND-MEAL: Improving Decision-Making through Data-Enabled Learning Concept Note*. Data-Pop Alliance. https://datapopalliance.org/wp-content/uploads/2018/01/Final_Concept-Note_Full-version-.pdf

de la Croix, D., Docquier, F., & Schiff, M. (2014). Brain Drain and Economic Performance in Small Island Developing States. In A. Artal-Tur, G. Peri, & F. Requena-Silvente (Eds), *The Socio-Economic Impact of Migration Flows: Effects on Trade, Remittances, Output, and the Labour Market* (pp. 123–144). Springer International Publishing. https://doi.org/10.1007/978-3-319-04078-3_6

Devarajan, S. (2011, October 6). *Africa's statistical tragedy*. World Bank Blogs. https://blogs.worldbank.org/africacan/africa-s-statistical-tragedy

Dickinson, E. (2011, January 3). GDP: A brief history. *Foreign Policy*. https://foreignpolicy.com/2011/01/03/gdp-a-brief-history/

The Economist. (2010, February 27). *The data deluge and how to handle it: A 14-page special report*. https://www.economist.com/weeklyedition/2010-02-27

Eurostat. (2022). *Empowering society by reusing privately-held data for official statistics—A European approach—2022 edition*. Eurostat. https://ec.europa.eu/eurostat/web/products-statistical-reports/-/ks-ft-22-004

Georgescu-Roegen, N. (1971). *The Entropy Law and the Economic Process*. Harvard University Press.

Georgescu-Roegen, N. (2020). *La Décroissance—Entropie—Ecologie—Economie*. SANG TERRE.

Giugale, M. (2013, December 18). *Fix Africa's Statistics | HuffPost The World Post*. HuffPost. https://www.huffpost.com/entry/fix-africas-statistics_b_2324936

Gnegne, Y. (2009). Adjusted net saving and welfare change. *Ecological Economics*, *68*(4), 1127–1139. https://doi.org/10.1016/j.ecolecon.2008.08.002

GPSDD. (2018). *Better Data. Better Decisions. Better Lives.* https://www.data4sdgs.org/sites/default/files/2018-03/About%20the%20Global%20Partnership_30Nov.pdf

Hellerstein, J. (2008, November 19). The Commoditization of Massive Data Analysis. *Radar - Insight, Analysis, and Research for Emerging Technologies*. http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html

Hille, P. (2022, February 23). *Democracy in decline worldwide*. Dw.Com. https://www.dw.com/en/democracy-in-decline-worldwide/a-60878855

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–721.

King, G. (2016, March 17). *Big Data is Not About the Data!* Informatics Symposium, University of Florida, Gainesville, United States. https://gking.harvard.edu/presentations/big-data-not-about-data-12

Kolbert, E. (2017, February 19). Why Facts Don't Change Our Minds. *The New Yorker*. https://www.newyorker.com/magazine/2017/02/27/why-facts-dont-change-our-minds

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lee, R. (2003). The demographic transition: Three centuries of fundamental change. *The Journal of Economic Perspectives*, *17*(4), 167–190. https://doi.org/10.1257/089533003772034943

Lepore, J. (2020, April 8). *The End of Knowledge: How Data Killed Facts*. https://scholar.harvard.edu/jlepore/presentations/emerging-technologies-and-future-humanities

Letouzé, E. (2012). *Big Data for Development: Opportunities and Challenges*. UN Global Pulse. https://www.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay2012.pdf

Letouzé, E. (2015). *Big Data and Development: An Overview* (Data-Pop Alliance Primers Series). Data-Pop Alliance. https://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf

Letouzé, E., Meier, P., & Vinck, P. (2013). Big Data for Conflict Prevention: New Oil and Old Fires. In F. Mancini (Ed.), *New Technology and the Prevention of Violence and Conflict* (pp. 4–27). International Peace Institute. https://www.ipinst.org/images/pdfs/IPI_Epub-New_Technology-final.pdf

Letouzé, E., & Oliver, N. (2019). *Sharing is Caring: Four Key Requirements for Sustainable Private Data Sharing and Use for Public Good*. Data-Pop Alliance. https://datapopalliance.org/wp-content/uploads/2019/11/DPA_VFI-SHARING-IS-CARING.pdf

Letouzé, E., Oliver, N., Lepri, B., & Vinck, P. (2023). AI for the SDGs—And Beyond? Towards a Human AI Culture for Development and Democracy. In B. Prud'homme, C. Régis, & G. Farnadi (Eds), *Missing Links in AI Governance* (pp. 159–189). United Nations Educational, Scientific and Cultural Organization (UNESCO). https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000384787&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_c75d935a-735f-435d-b54b-2a288c57da69%3F_%3D384787eng.pdf&locale=en&multi=true&ark=/ark:/48223/pf0000384787/PDF/384787eng.pdf#A14974_Mila_MissingLinksAI_8x10_AN_EP10.indd%3A.142539%3A1147

Letouzé, E., & Pentland, A. (2018). Human AI for Human Development. *ITU Journal: ICT Discoveries*, *1*(2). https://www.itu.int/en/journal/002/Pages/15.aspx

Letouzé, E., & Yáñez Soria, I. (2021, August 3). *The CODE for Building Participatory and Ethical Data Projects*. Data-Pop Alliance. https://datapopalliance.org/the-code-for-building-participatory-and-ethical-data-projects/

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

McArthur, J., & Rasmussen, K. (2017, January 11). Change of pace: Accelerations and advances during the Millennium Development Goal era. *Brookings*. https://www.brookings.edu/articles/change-of-pace-accelerations-and-advances-during-the-millennium-development-goal-era/

Mercier, C., & Aubert, S. (2023, July 7). *Michel Poivert, historien d'art: Photographe, «un antihéros de l'information»*. Libération. https://www.liberation.fr/culture/

photographie/michel-poivert-historien-dart-photographe-un-antiheros-de-linformation-20230707_EFNEUZKAOJGJFEYI2WXC2WZLRY/

Noor, A. M., Alegana, V. A., Gething, P. W., Tatem, A. J., & Snow, R. W. (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics*, *6*(1), 5. https://doi.org/10.1186/1478-7954-6-5

OHCHR. (2014, November 20). *Better data, better lives*. OHCHR. https://www.ohchr.org/en/stories/2015/10/better-data-better-lives

Papageorgiou, C., & Johnson, P. (2019, April 16). *It's too soon for optimism about convergence*. CEPR. https://cepr.org/voxeu/columns/its-too-soon-optimism-about-convergence

Porway, J. (2021, July 14). The Three Uses of Data. *Data.Org*. https://data.org/news/the-three-uses-of-data/

Przeworski, A. (2004). Institutions matter? *Government and Opposition*, *39*(4), 527–540. https://doi.org/10.1111/j.1477-7053.2004.00134.x

Riley, J. C. (2005). Estimates of regional and global life expectancy, 1800–2001. *Population and Development Review*, *31*(3), 537–543.

Roser, M., Ortiz-Ospina, E., & Ritchie, H. (2013). Life Expectancy. *Our World in Data*. https://ourworldindata.org/life-expectancy

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press.

Stuart, E., Samman, E., Avis, W., & Berliner, T. (2015). *The data revolution: Finding the missing millions* (3; Research Report). Overseas Development Institute. https://cdn.odi.org/media/documents/9604.pdf

TSIA (Director). (2013, November 9). *The New Data Refineries: Transforming Big Data Into Decisions - Andreas Weigend*. https://www.youtube.com/watch?v=H4IFQUQWjVE

UNDP. (2023a). *Background on the goals*. United Nations Development Programme. https://www.undp.org/sdg-accelerator/background-goals

UNDP. (2023b). *Human Development Index (HDI)*. United Nations Development Programme. https://hdr.undp.org/data-center/human-development-index

United Nations. (2013). *A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development*. United Nations. https://www.post2020hlp.org/the-report/

United Nations. (2017). *SDG Indicators: Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development*. United Nations. https://unstats.un.org/sdgs/indicators/indicators-list/

United Nations. (2022a). *IAEG-SDGs: The Classification for Global SDG Indicators*. Sustainable Development Goals. https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/

United Nations. (2022b). *SIDS Accelerated Modalities of Action [S.A.M.O.A.] Pathway*. Sustainable Development Knowledge Platform. https://sustainabledevelopment.un.org/sids2014/samoapathway

United Nations Secretary-General. (2013). *Report of the Secretary-General's High-Level Panel of Eminent Persons on the Post-2015 Development Agenda*. United Nations Secretary-General. https://www.un.org/sg/en/management/beyond2015.shtml

Vincent, J. (2022). *Beyond Measure: The Hidden History of Measurement from Cubits to Quantum Constants*. W. W. Norton & Company.

West, E. G. (1978). Literacy and the industrial revolution. *The Economic History Review*, *31*(3), 369–383. https://doi.org/10.2307/2598759

World Bank. (2023). *Statistical Performance Indicators (SPI)*. World Bank. https://www.worldbank.org/en/programs/statistical-performance-indicators

PART II

THE EMERGING USES OF BIG DATA ANALYSIS IN HUMAN DEVELOPMENT

Case Studies from Around the World

5

Predicting Socioeconomic Status Using Big Data

Christopher Njuguna and Patrick McSharry

5.1 Introduction

Less than 10% of the world's population or approximately 605 million people are currently considered extremely poor (World Data Lab, 2023). According to the World Bank, the COVID-19 pandemic, rising inflation and war in Ukraine is estimated to have led to an additional 75 million to 95 million people living in extreme poverty in 2022, compared to pre-pandemic projections. While there has been a significant decline in extreme poverty from a global estimate of 36% (approximately 1.9 billion) in 1990, the number of global poor still remains a major challenge. The UN General Assembly, in adopting the agenda for sustainable development, stated that ending poverty in all its forms and dimensions, including extreme poverty, is "the greatest global challenge and an indispensable requirement for sustainable development" (UN, 2015). It then follows that the prime sustainable development goal, SDG1, pledges to "end poverty in all its forms everywhere" with the target "to eliminate extreme poverty for all people everywhere by 2030." To eliminate poverty, it must first be defined and measured. For this we use poverty indices which fall into two main categories: monetary and non-monetary poverty indices (Bourguignon and Chakravarty, 2003). Monetary indices are based on monetary income measures and include national poverty lines and the World Bank poverty indices. Non-monetary indices rely on proxies for wealth and assign poverty level by how much one is deprived of the listed proxies. One such non-monetary index is the Multi-Dimensional Poverty Index (MPI) developed at the University of Oxford (Oxford Poverty and Human Development Initiative, 2013). The UN and World Bank revised the extreme poverty level from $1.25 to $1.90 a day in 2015 and most recently to $2.25 in 2022.

Christopher Njuguna and Patrick McSharry, *Predicting Socioeconomic Status Using Big Data*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press.
© Christopher Njuguna and Patrick McSharry (2025). DOI: 10.1093/oso/9780198879145.003.0006

Poverty indices are among the key figures and indicators that are used to influence policy and, thus, need to be based on data that are representative of the population. Traditionally, the census and surveys are two tools that have been commonly used to obtain such data. While these tools have been proven to be accurate and reliable, they have high costs—monetarily, in time, and in effort—and can therefore only be undertaken periodically, typically in three-to-ten-year cycles (Blumenstock et al., 2015) (see Table 5.1 for censuses and surveys regularly organized in Rwanda). The census presents a one-off detailed socioeconomic snapshot of a country or region that is undertaken every ten years but only allows for the assessment of policies after three to ten years. Furthermore, census data take time to be analyzed and disseminated and are literally out of date by the time of use. Worse still, in some developing countries, especially in Sub-Saharan Africa, there has not been an official census or survey in decades, leading to a "dearth of reliable statistics"—a challenge that the World Bank terms Africa's "statistical tragedy" (Letouzé, 2014). This lack of data and their low update frequency has prompted a global search for alternatives to these traditional tools.

The digitization of records and the proliferation of mobile phones and remote sensing devices have led to an explosion in the amounts and variety of available data sources. This explosion of data has led to a phenomenon known as "big data" which is showing promise as an alternative to the traditional data sourcing tools (Global Pulse, 2013). Big data has the advantage that it provides a "highly mobile, location-aware, person-centered and context-relevant" (Chen et al., 2012) dataset. Big data is causing a revolution in both the public and private sectors and innovative predictive analytics

Table 5.1 Censuses and surveys in Rwanda as of December 2015.

Name	Frequency (years)	Last year	Households sampled
Population and Housing Census (RPHC4)	10	2012	Whole population
Housing and Living Conditions Survey (EICV4)	3	2013/14	14,419
Demographic and Health Survey (DHS5)	3	2014/15	12,793

is leading to the creation of new intellectual property with many applications (Thomas & McSharry, 2015). The UN has recognized the positive impact of big data in projects and called for its increased use to inform decision-making and to assist in achieving the SDGs, terming it the "data revolution for development." In line with this new direction, the UN Global Pulse was created as a "flagship innovation initiative of the United Nations Secretary-General on big data" based on the recognition that "digital data offers the opportunity to gain a better understanding of changes in human well-being and to get real-time feedback on how well policy responses are working" (Global Pulse, 2013). This activity within the UN demonstrates that big data is not only of academic interest but is already being considered for policymaking.

Big data, however, has its limitations. One of the characteristics of big data is that it is not usually collected with the intent to analyze it (Letouzé, 2014). This means that the sample of the data collected may not be representative of the population due to various biases (Hilbert, 2013). Second, due to the cost and capacity required to store and process big data, the developing world may not benefit from it as much as the more developed world, widening the digital divide (Letouzé, 2014). A third limitation is in the disproportionate power given to certain actors in the big data ecosystem. For example, the data analyst chooses the data to collect and the methods they use to analyze and present them (Taylor et al., 2014). Finally, big data has limitations in access due to privacy and commercial concerns (Hilbert, 2013). The privacy issue may be overcome by aggregating datasets, however, merging datasets, even aggregated ones, may have the effect of revealing even anonymized individual data (Taylor et al., 2014). Another way to overcome "closed" data is to showcase benefits achieved in other national contexts (Taylor et al., 2014). As big data grows in influence and as more uses are found for it, it will continue to be a crucial subject area for researchers and practitioners ahead of mainstream acceptance.

It has been found that researchers can combine big data with other data sources to yield high-resolution datasets which can be used in analysis and prediction when compared to a ground truth variable such as census data (Taylor et al., 2014). One form of big data that is gaining attention from researchers is the phone record collected by mobile companies for billing purposes. Mobile phone records, otherwise called call detail records (CDRs), are continuously available at a daily frequency and high spatial resolution associated with individual cell towers. Such CDR data could be available to government at a reduced cost since it is passively collected and can, therefore, be viewed as a social good.

However, mobile operators are reluctant to release CDRs to the public domain due to commercial sensitivities. There is also a growing fear of privacy risk whereby anonymized data could be de-anonymized using sophisticated techniques (Letouzé et al., 2015). Sparse CDRs containing limited fields in combination with other publicly available datasets may provide a reasonable balance between risk and opportunity and pave the way for a workable solution. We find that the use of such a pared down CDR dataset in conjunction with such datasets as satellite night lights can result in similar or better estimates of socioeconomic indicators than those derived using richer CDR datasets alone, for example.

5.2 Background: Literature Review

Blumenstock & Eagle (2010) carried out a study in Rwanda to determine the socioeconomic status of individuals based on their CDRs. They were able to predict asset ownership with a correlation coefficient of 0.917 to the 2007 DHS3 asset ownership data. Blumenstock et al. (2015) extended their original research, estimating subscriber socioeconomic status with a cross-validated correlation coefficient r = 0.68 at the cell and smaller spatial resolutions. Smith-Clarke et al. (2014) determined that regions with fewer and smaller volumes of calls with other regions were more likely to have a lower socioeconomic status. Soto et al. (2011) were able to determine the socioeconomic levels of regions with a classification accuracy of up to 80%.

Satellite nightlight data have been used to determine economic activity at national and sub-national levels. For example, in 2009, Elvidge et al. (2009) built a poverty index based on nightlight data and Landscan population data, estimating a global total of 2.2 billion people below the poverty line against World Bank estimates of 2.6 billion—an error of 15%. In another study, Mellander et al. (2015) found that nightlight-based features, especially light density, are a good proxy for economic activity.

An overview of some of these studies and the sources of data are provided in Table 5.2. In this study, we demonstrate that just a few carefully chosen CDR features combined with other publicly available datasets are sufficient to accurately determine the poverty level of a sub-national spatial region. This is important because researchers often only have access to limited datasets due to commercial and privacy concerns. The successful use of this restricted and anonymized dataset to determine the poverty levels of a region could offer considerable opportunities to enhance the work of policymakers, donors and non-governmental organizations (NGOs).

Table 5.2 Chronological sample of related studies and a description of the sources of data.

Author	Data source	Data year	Country/ region/ individual	Population sample size	Time period	# Records	Poverty measure	Model	Correlation
C. Elvidge et al (2009)	Nightlight	2009	Global & national	–	12 months	–	WDI[a] $2 a day international poverty line (2006)	Regression	0.85
Blumenstock & Eagle (2010)	CDR & Survey	2005–2008 & 2009	Individual	1.5M & 856	9 months	1B+ & 64K	–	–	–
Soto et al (2011)	CDR	2010	Main city in Latin American country	500K	6 months	Unknown	Socioeconomic levels (A, B, C) from NSI[b]	Support Vector Machine	0.80
Smith-Clarke et al (2014)	CDR	2011/12 and 2012	Cote d'Ivoire and Anonymous Region B	5M & 928K	20 weeks & 6 weeks	471M and 40M	IMF poverty rate estimates (2008)	OLS[c] regression	–
Blumenstock et al (2015)	CDR & survey	2009	Rwanda	1.5M & 856	9 months	1B+ and 64K	DHSIV Composite Wealth Index	Linear regression*	0.68
This study (2015)	CDR, RPHC4, Landscan and nightlight	2014/15	Rwanda	4.8M	9 months	3.8B	RPHC4 MPI	Linear regression*	0.88

[a] World Development Index.
[b] National Statistics Institute.
[c] Ordinary Least Squares regression.
* Cross-validated.

5.3 Methodology

Unlike other studies that use automated methods (e.g., Blumenstock and Eagle, 2012; Blumenstock et al., 2015; Smith-Clarke et al., 2014), we employ manual feature engineering, analysis and selection of the features in the final model. An overview of the procedure follows.

After obtaining the data, a small number ($n < 20$) of intuitive features were constructed with the aim of offering proxies for economic activity or socioeconomic status. The data were aggregated at the sector level to match the spatial resolution of the ground truth given by the MPI. In line with our objective to build a self-updating index, we left out features that had low update frequencies (e.g., population from census data). Any features that had low correlation to the MPI ($r < 0.4$) were also dropped.

We noted that the actual MPI is approximately normally distributed while all our predictors were highly skewed. These variables were normalized to reduce the effects of population and geographical area.

Feature selection was achieved using (i) LASSO (Least Absolute Shrinkage and Selection Operator) to eliminate any features that were not relevant for constructing the model, (ii) ridge regression to eliminate any collinear features, and (iii) elastic nets to ensure the models were sparse. To obtain significance values for each feature, and as a farther confirmation of the features selected for the final model, backwards stepwise regression was used.

In the next sections we describe the different datasets used and how they were processed and transformed to form the features used to build the final model.

5.3.1 Data

5.3.1.1 Sectors

The spatial component for this study is the sector in Rwanda, which is the third administrative level after the province and district. There are 416 sectors, 30 districts and 5 provinces in Rwanda. Sector maps were obtained from the Rwanda Land Use Planning Portal ("Rwanda Land Use Planning Portal," n.d.). Sectors in Rwanda have areas ranging from 1 km^2 to 650 km^2 and an average of 58 km^2.

5.3.1.2 The Multi-dimensional Poverty Index (MPI)

The multi-dimensional poverty index (MPI) is an index that was developed at the University of Oxford's Department of International Development

under the Oxford Poverty and Human Development Initiative (OPHI, 2013). The MPI measures three dimensions: health, education, and standard of living (Oxford Poverty and Human Development Initiative, 2013). The MPI data used in this study were based on data from the 2012 Rwanda Population and Housing Census (RPHC4) and are representative of the levels of poverty per sector in the country.

5.3.1.3 Call Detail Records (CDRs)

We use mobile CDRs from a leading mobile network operator (MNO), with about 49% market share (as of the last month in the dataset) in Rwanda, representing calls and SMSs between July 1, 2014 and March 31, 2015. In this dataset, user IDs are anonymized, calls are not distinguished from SMSs, do not contain call duration, and do not indicate the receiving party. The call records contained the following fields: timestamp, anonymized subscriber ID, start cell and end cell. Summary statistics of the CDRs are provided in Table 5.3.

The CDRs are mapped to 580 towers located across the country as shown in Figure 5.1, whose locations are used to estimate the subscriber location at the time of the event.

In this study we focus on two variables that are extracted from the CDRs: mobile ownership and call volume.

5.3.1.4 Nightlight Satellite Data

Nightlight data (Earth Observation Group, n.d.) are records of artificial light on the earth's surface collected by satellites. The nightlight maps have a resolution of 15 arc-seconds which corresponds to about 450m x 450m at the equator and have a range of 3 x 10^{-9} W.cm^{-2}.sr^{-1} to 0.02 W.cm^{-2}.sr^{-1} (Shi et al., 2014). The data used is a monthly composite of daily nightlight intensities for the month of December 2014.

5.3.1.5 Landscan Population Data

The Landscan population database (Rose and Bright, 2014) produced by the US Department of Energy, Oak Ridge National Library, is a spatially disaggregated global population count dataset. Unlike census data which represents residential populations, Landscan represents a 24-hour

Table 5.3 Statistics from the Rwanda CDR Dataset: July 2014–March 2015.

Number of calls/SMSs	Number of subscribers	Number of cells	Number of towers
3.8B	4.8M	3075	580

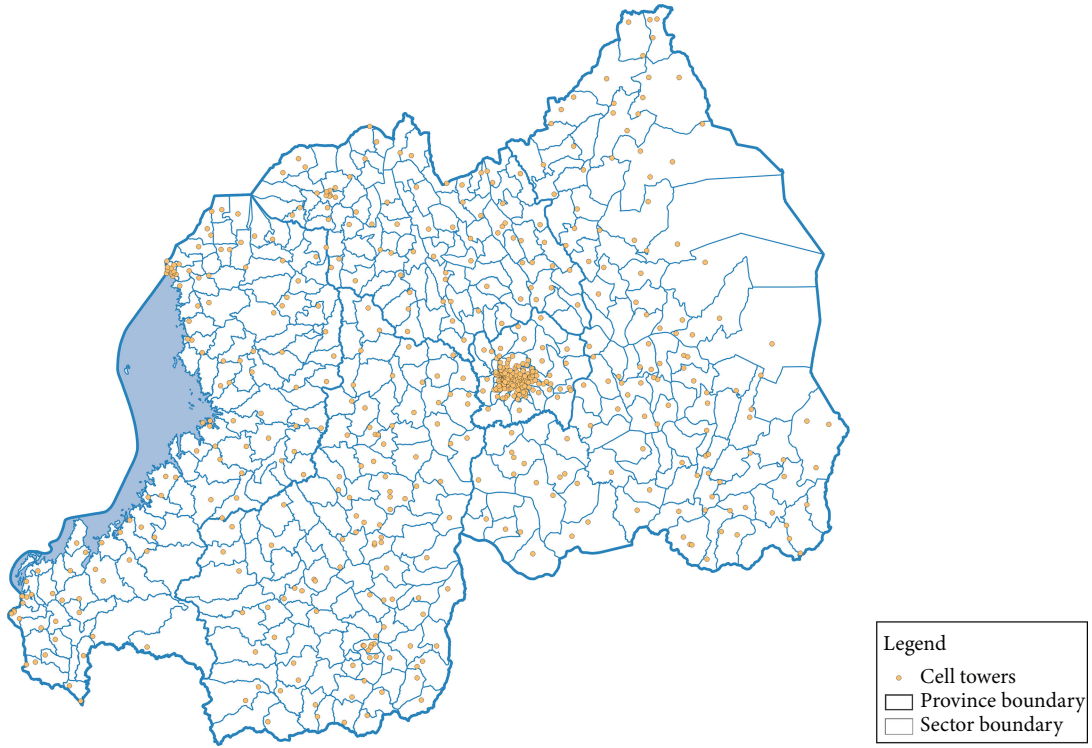


Figure 5.1 Rwanda sector map showing cell tower locations.

population (Rose and Bright, 2014). We use this dataset to show that poverty levels can be calculated using wholly independent and unofficial sources. We found a correlation of 0.87 between the 2012 Landscan and the 2012 census (RPHC4) population counts for Rwanda at the sector level.

5.3.2 Feature Engineering

To determine the socioeconomic status of the sectors in Rwanda, we calculated a small number of meaningful features from the CDRs combined with the nightlights and Landscan population datasets. We now provide a detailed description of the construction of each of the selected features.

5.3.2.1 Nightlight per Capita

While nightlights are used as a proxy for economic activity, the amount of activity that can be attributed per capita yields a socioeconomic profile for the individual and hence of the area. Higher intensities with lower populations would yield the highest socioeconomic status of a region while low nightlight intensities of areas with high populations would imply a lower socioeconomic status. We divide the total intensity in each sector by the population of the sector to yield a light intensity per capita. The nightlight intensity per capita has a correlation of 0.68 with the MPI. The log–log relationship indicates that, with other factors held constant, the MPI is expected to reduce by—0.12% for every 1% increase in nightlight intensity per capita.

5.3.2.2 Average Daily Call Volume per Phone

It has been shown that the higher the socioeconomic level, the more calls a subscriber would make on a daily basis (Blumenstock and Eagle, 2012). Call volume per phone is calculated as the total number of events (calls and SMSs) made from the mobiles owned in a particular sector and averaged to provide a daily call volume. Dividing this number by the number of days in the dataset yields the average daily call volume per phone. The average daily call volume per phone has a correlation of 0.44 with the log of MPI. The log-linear relationship indicates that the MPI decreases by—2.16% for an increase of one call per day.

5.3.2.3 Mobile Ownership per Capita

Mobile phone ownership is a good indicator of socioeconomic status. For example, 85% of households in the Province of Kigali owned a mobile

phone compared to only 46% in the Southern Province (National Institute of Statistics of Rwanda (NISR) and (MINECOFIN), 2014). CDR data are not fully representative of the population since the poorest and most vulnerable sections of society may not be represented. Mobile ownership is derived from CDRs by determining unique mobile identities and the towers at which they are most active, then aggregating them to the sector in which the towers are located. Note that sectors without towers are not included in this model. Mobile ownership per capita has a correlation of 0.47 to the log of MPI. The log-linear relationship suggests that the MPI will decrease by 14.25% when the number of mobile subscribers per capita increases by one unit.

5.3.2.4 Population Density

Global trends show that urbanization is increasing, and this provides an indicator of industrialization. It is assumed that the greater the population density, the higher the probability of economic activity. To calculate the population density, we use the sector population data from the Landscan database and divide it by the sector area. The log of the population density has a correlation of 0.8 with the log of MPI, and the log–log relationship implies that the MPI decreases by 0.32% when the population density increases by 1%.

5.4 Results

The predictive features are carefully constructed so as to provide proxies for economic activity in each sector. For this reason, we have to ensure that each feature is normalized such that neither the population nor area of the sector influence the measured values. In addition, we need to transform each feature so that it is approximately normally distributed to facilitate the construction of a quantitative model.

Analysis of the scatter plots of the four features against the MPI suggested possibly non-linear relationships between nightlights per capita and population density and MPI. To attempt to fit a linear model to this data, we transformed the data by taking their logarithms. To assess the validity of this approach, we used LASSO feature selection to decide whether the log or linear versions of the variables should be employed. The linear versions of the call volume per phone and the number of phones per capita were selected. Table 5.4 provides summary statistics for the four predictive features and the MPI. The scatter plots showing the transformed relationships are provided below in Figure 5.2.

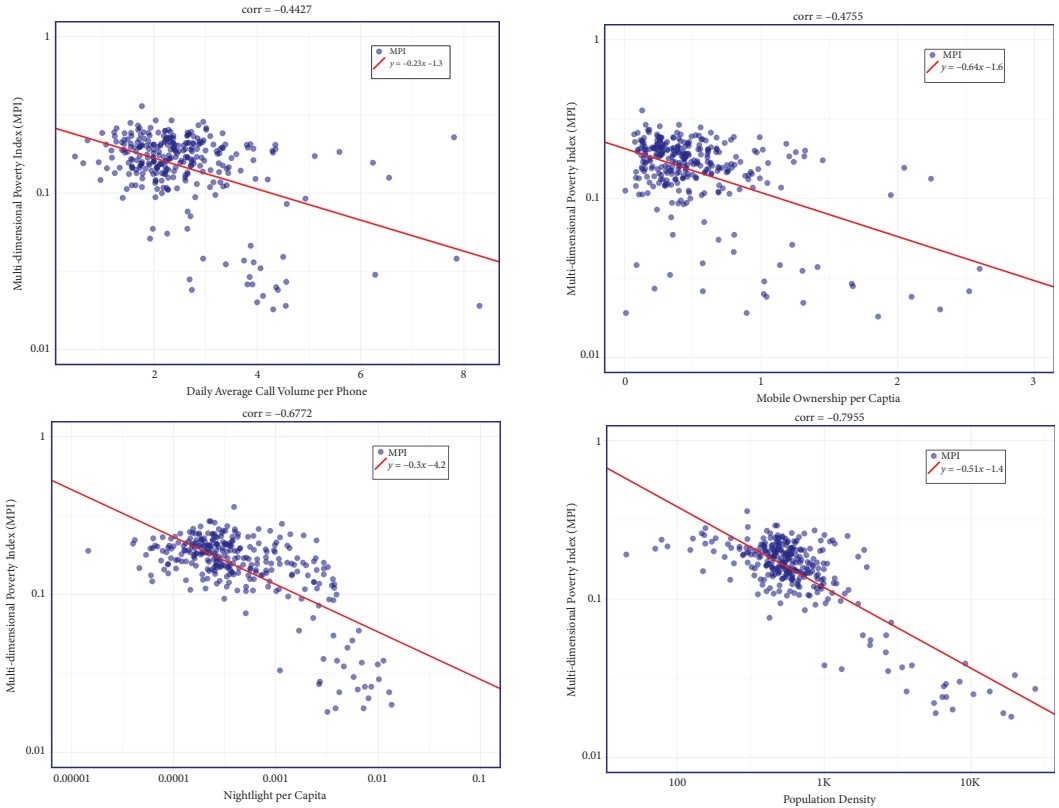


Figure 5.2 Scatter plots of the MPI vs the four predictive features showing their correlations.

Table 5.4 Summary statistics of the features and MPI.

Feature	Average	Median	Minimum	Maximum
Nightlights per capita	0.00008	0.00028	0.00001	0.01328
Daily call volume per phone	2.54	2.30	0.48	8.30
Mobile ownership per capita	0.523	0.413	0.007	2.601
Population density	1032	551	45	27333
MPI	0.1703	0.1730	0.0180	0.3590

Using LASSO, we fit a model between the four features and the MPI, yielding a linear model with the following structure:

$$log\,log\,(MPI) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (1)$$

The parameter estimates are given in Table 5.5 and were all statistically significant ($p < 0.05$). A comparison of the predictability of the individual features and the final model is provided in Table 5.6. The cross-validated correlation coefficient of our predicted log (MPI) versus the actual log (MPI) is 0.88. This corresponds to an R-Squared coefficient of 0.76 and means that the selected model explains 76% of the variance of the MPI across the 295 sectors in Rwanda. Figure 5.3 shows the predicted MPI versus the actual MPI. We can then rank the sectors in deciles in terms of their poverty level to determine the sectors with the highest and lowest poverty levels in Rwanda. This information can then be used to construct a visual graphic in the form of a map as shown in Figure 5.4.

The maps in Figure 5.4 show the decile ranking of sectors based on their MPIs, the first being the MPI obtained from the NISR and the second the

Table 5.5 Estimated parameter values of the multivariate linear model consisting of four predictive features.

i	x_i	β_i
0	Intercept	–0.5924
1	Log (nightlights per capita)	–0.1187
2	Average Daily Call volume per phone	–0.0218
3	Mobile ownership per capita	–0.1537
4	Log (Population density)	–0.3244

Table 5.6 Correlations of the four predictive features and the final model prediction with the MPI.

Feature	Correlation	t-stat	p-value
Log (nightlights per capita)	−0.6772	−9.284	2.09×10^{-18}***
Call volume per phone	−0.4427	−3.570	4.76×10^{-4}***
Mobile ownership per capita	−0.4764	−5.982	5.51×10^{-9}***
Log (Population density)	−0.7955	−15.523	4.42×10^{-40}***
Model	0.8764		

* significant at p = 0.05.
** significant at p = 0.01.
*** significant at p = 0.001.

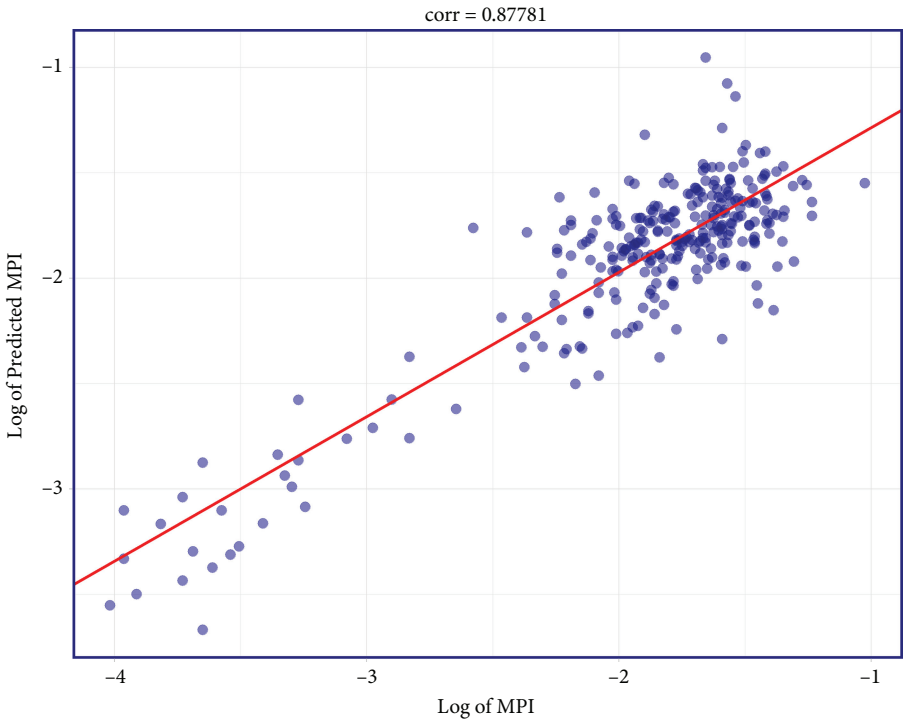


Figure 5.3 Scatter plot of the predicted MPI vs the actual MPI.

estimated MPI from our model. The estimated MPI map is created using data imputed using the nearest neighbor technique for the 121 sectors that do not have cell towers in order to aid visual analysis. The correlation in the maps is most visible where there are clusters of sectors that fall in the same or neighboring deciles. This can be seen around Kigali (in the center) where

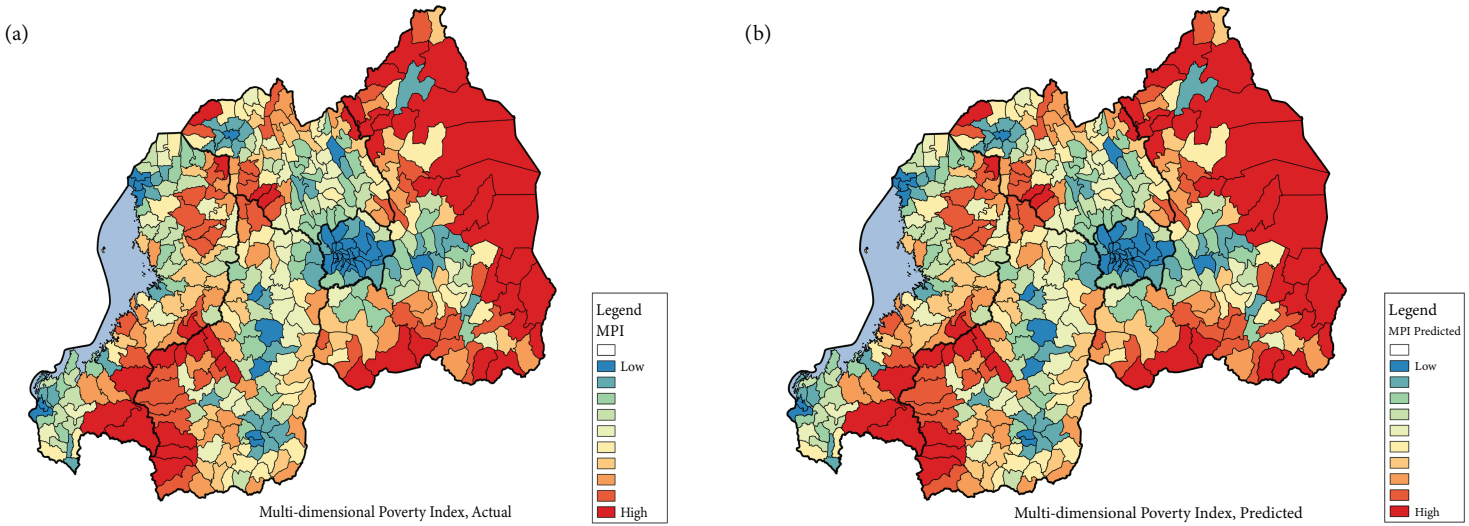


Figure 5.4 Maps showing the actual (a) and estimated (b) MPI in deciles.

there is low MPI, and in clusters to the south, southwest and northwest, where the MPI is high.

5.5 Discussion

5.5.1 Limitations

This study utilizes CDRs obtained from a leading MNO with a greater market share in urban areas. However, CDRs from multiple MNOs would eliminate biases introduced by their positioning and marketing campaigns and provide statistics that are more representative of the various socioeconomic levels in Rwanda. Second, only 295 of the 416 sectors contain mobile phone cell towers. This means that 121 sectors do not have a direct measure of CDR statistics. We realize that this means that there is a definite bias introduced for the sectors where data is available, and estimates are calculated. Finally, there is a lag between the predictor data and the ground truth data, i.e., the MPI. While the nightlights, Landscan population and CDRs are all gathered from periods around the year 2014, the MPI is calculated using data from the 2012 Rwanda population census. This lag is likely to have negatively affected our results. Using datasets picked from the same time period should yield even better results.

5.5.2 Future Work

This study utilizes CDRs that are located only in the sectors where cell towers exist. Other studies have used various techniques such as Voronoi partitions to distribute the tower-based statistics, e.g., call volume and mobile ownership, to finer geographical resolutions in the country via an intersection and merge operation. It is envisaged that, with some additional analysis in this direction, suitable features that are representative of economic and socioeconomic status for arbitrary geographical areas can be estimated.

Secondly, while this study focuses on sub-country regions, this approach can be applied to multi-country contexts where comparison across countries may be of interest. With data collected for different countries, and using the corresponding country MPI ground truths, it would be possible to build a country-level global or regional poverty index.

One of the concerns raised in this study is the availability and use of CDRs from only one MNO which, may not necessarily be representative of the entire population. This is a major concern and one that gets compounded as larger regions such as the African continent are taken into consideration. Different MNOs will have differing market shares in different countries and regions. In order to build a model that is representative of the population, CDRs would ideally be collected from all MNOs thereby providing full coverage of the population. Even then, it may be necessary to consider the various factors that affect model features, such as mobile phone ownership and call volume. These may include mobile access, MNO pricing strategies, government subsidies and sanctions, as well as forex rates.

Finally, future work should result in the building of a self-updating system with pipelines from all data sources and all processing automated where possible. An example frontend would be a dynamic web-based country map showing the sectors ranked and color coded based on the estimated MPI, which could have a playback function to show how the ranking has changed over time.

5.6 Conclusion

In contrast to censuses and surveys which are resource intensive and carried out in cycles spanning multiple years, big data has been shown to be an effective means of obtaining socioeconomic estimates which can be used to continuously monitor the effects of policy and to help shape future decisions. We have demonstrated that the use of independent, publicly available and relatively cheap data sources such as call detail records, nightlights, and the Landscan population dataset can facilitate accurate estimation of the multidimensional poverty levels of the sectors in Rwanda. We fit a multivariate linear model that describes the relationship between four features, nightlight intensity per capita, call volume per phone, mobile ownership per capita and population density to the multi-dimensional poverty index. We estimate MPIs for 295 of the 416 sectors in Rwanda that have mobile phone towers within their boundaries. The key findings of the research are the following:

- Demonstrating the predictability of poverty levels based on proxies for disposable income such as nightlights and mobile telephone ownership and usage.

- Showing that a sparse subset of call detail records, mitigating commercial and privacy risks, can be used to accurately quantify poverty.
- Constructing a model based on big data sources that allows for accurate prediction of the MPI with a cross-validated correlation coefficient of r=0.88.
- Establishing the potential of big data for informing policymakers and monitoring poverty in real time by geographical region.
- Providing the basis of an approach for government, donors, NGOs, and the private sector to target resources and monitor the impact of different interventions.
- Paving the way for business applications to assess the potential of providing banking services for the poorest of the poor, improve financial inclusion and facilitate credit risk ratings.

Importantly, this study yields competitive results from a sparse CDR dataset, avoiding fields that MNOs consider commercially sensitive or privacy risks. We believe that this may provide a compromise between the concerns and motivations of MNOs and policymakers in the quest to unlock the wealth of data that CDRs represent.

Finally, we show that big data is a valuable resource for estimating socioeconomic indicators and should be accepted as a dependable and accurate source of information. Timeliness and high temporal resolution are two enormous advantages of big data that should be emphasized. The minimal loss of accuracy in predicting MPI is more than made up for by the near real-time aspect of this innovative policy tool based on big data.

Acknowledgments

We would like to thank Patrick Nyirishema, the Director General of Rwanda Utilities Regulatory Authority (RURA), for partnering with Carnegie Mellon University on this project. We would also like to thank Francis Ngabo, Georges Kwizera, Protais Kanyankore and the staff of RURA in general for their support. We are grateful to Rajiv Ranjan, Governance Unit of the United Nations Development Program (UNDP) and ICT Advisor to the National Institute of Statistics of Rwanda (NISR) and Tom Bundervoet, Senior Economist with the World Bank for their invaluable help and support. We acknowledge Chuma Vuningoma and the staff of MTN, Rwanda for their support and assistance with CDR acquisition.

References

Blumenstock, J. E., Cadamuro, G., and On, R. 2015. *Predicting Poverty and Wealth from Mobile Phone Metadata.* Working Paper. 350. https://doi.org/10.1126/science.aac4420

Blumenstock, J. E., and Eagle, N. 2012. Divided we call : Disparities in access and use of mobile phones in Rwanda. Information Technology and International Development Journal, Volume 8, 1–16.

Blumenstock, J., & Eagle, N. (2010). Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, 1–10. https://doi.org/10.1145/2369220.2369225 DOI: https://doi.org/10.1145/2369220.2369225

Bourguignon, F., and Chakravarty, S., 2003. The measurement of multidimensional poverty. The Journal of Economic Inequality, Vol. 1, Issue 1, 25–49. https://doi.org/10.1007/978-0-387-79253-8

Chen, H., Chiang, R. H. L., and Storey, V. C., 2012. Business intelligence and analytics: From big data to big impact. Management Information Systems Quarterly Special Issue on Business Intelligence. 36, 1165–1188.

Earth Observation Group, N., n.d. Earth Observation Group Suite of Average Radiance Composite Images using Nighttime Data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) [WWW Document]. URL https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html (accessed 8.12.15).

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., and Bright, E. 2009. A global poverty map derived from satellite data. Computers & Geosciences, Volume 35, Issue 8 https://doi.org/10.1016/j.cageo.2009.01.009

Global Pulse. 2013. Mobile phone network data for development. United Nations Global Pulse 1–12.

Hilbert, M. 2013. *Big Data for Development: From Information to Knowledge Societies*. http://dx.doi.org/10.2139/ssrn.2205145

Letouzé, E. 2014. Big Data for Development: Facts and Figures, Institutions *7*, 10.

Letouzé, E., Vinck, P., and Kammourieh, L. 2015. *The Law, Politics and Ethics of Cell Phone Data Analytics*. Data-Pop Alliance White Paper Series.

Mellander, C., Lobo, J., Stolarick, K., and Matheson, Z. 2015. Night-time light data: A good proxy measure for economic activity? *PLoS One* 10, 1–18. https://doi.org/10.1371/journal.pone.0139779

National Institute of Statistics of Rwanda (NISR), (MINECOFIN), Ministry of Finance and Economic Planning 2014. Fourth Population and Housing Census, Rwanda, 2012. Thematic Report: Characteristics of Households and Housing 9–19.

Oxford Poverty and Human Development Initiative. 2013. *Rwanda Country Briefing. Multidimens. Poverty Index Data Bank*. OPHI, University of Oxford.

Rose, A.N., and Bright, E., 2014. The LandScan Global Population Distribution Project: Current State of the Art and Prospective Innovation.

Rwanda Land Use Planning Portal [WWW Document], n.d. URL http://www.rwandalanduse.rnra.rw (accessed 12.8.15).

Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., Chen, L., and Wu, J. 2014. Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data. Remote Sensing. Volume 6, Issue 2, 1705–1724. https://doi.org/10.3390/rs6021705

Smith-Clarke, C., Mashhadi, A., and Capra, L. 2014. Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks. In Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems—CHI '14. Association for Computing Machinery, New York, NY, USA. pp. 511–520. https://doi.org/10.1145/2556288.2557358

Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E. (2011). Prediction of Socioeconomic Levels Using Cell Phone Records. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds) User Modeling, Adaption and Personalization. UMAP 2011. Lecture Notes in Computer Science, vol 6787. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22362-4_35

Taylor, L., Cowls, J., Schroeder, R., and Meyer, E. T., 2014. Big data and positive change in the developing world. *Policy & Internet* 6, 418–444. https://doi.org/10.1002/1944-2866.POI378

Thomas R., McSharry P., Big Data Revolution: What farmers, doctors and insurance agents teach us about discovering big data patterns, John Wiley & Sons, 2015.

Transforming Our World: The 2030 Agenda for Sustainable Development [WWW Document], 2015. URL http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

United Nations. 2015. The Millennium Development Goals Report 2015. Department of Economics, United Nations Department of Public Information. 75.

World Data Lab.(2023. World poverty clock: Real time poverty forecasts. www.worldpoverty.io. (accessed 22.02.2023).

6

Reinventing the Social Contract in an Age of Big Data

Bruno Lepri, Nuria Oliver, and Alex Pentland

6.1 Introduction

Since the beginning of time humans have lived in a world where distance clouds perception, and where the experiences and the thoughts of people even just a few miles away are unknown. The spread of literacy, and later the access to rich media content such as photographs, recordings, and video, have enabled us to share experiences with a broader set of people, but have provided only a small window on reality, without much context. In fact, as recently as a few decades ago, it was possible to have a large-scale genocide (i.e., the Tutsi genocide in Rwanda), mass starvation (i.e., the 1983–1985 famine in Ethiopia), ecological disasters, and pandemics without the rest of the world knowing.

However, the wide adoption of Internet access, earth-imaging satellites, mobile phones with cameras, social media and similar technologies has dramatically changed the situation: now there is data about the world and about each of us everywhere, such that it is possible to imagine a world where each person counts—and where we cannot say that we didn't know that something happened somewhere, even if it's miles away from where we are. Recently, research studies have shown that this exponential increase in the volume and types of available data is opening unprecedented possibilities for understanding global patterns of human behavior and for helping decision makers to tackle problems of societal importance, such as monitoring socioeconomic deprivation (Soto et al., 2011; Smith-Clarke et al., 2014; Blumenstock et al., 2016; Jean et al., 2016; Steele et al., 2017) and crime levels (Bogomolov et al., 2014; Traunmueller et al., 2014; Wang et al., 2016), mapping the propagation of diseases (Wesolowski et al., 2012; Bengtsson et al., 2015; Wesolowski, Metcalf, et al., 2015; Wesolowski, Qureshi, et al., 2015; Ruktanonchai et al., 2016),

Bruno Lepri, Nuria Oliver, and Alex Pentland, *Reinventing the Social Contract in an Age of Big Data*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Bruno Lepri, Nuria Oliver, and Alex Pentland (2025). DOI: 10.1093/oso/9780198879145.003.0007

and understanding the impact of natural disasters, environmental risks, and other emergencies (Lu et al., 2012; Pastor-Escuerdo et al., 2014; Gundogdu et al., 2016; Lu et al., 2016; Wilson et al., 2016), etc. Thus, researchers, companies, governments, financial institutions, non-governmental organizations and also citizen groups are actively experimenting, innovating and adapting to the new world of data, a world in which data is bigger, faster and more detailed than ever before. However, while some of us are already living in this new world, a large fraction of the population, organizations and governments are excluded because of lack of resources, knowledge, capacity or opportunity.

For this reason, as the world embarks on the ambitious project to meet 17 Sustainable Development Goals (SDGs) by 2030, including no poverty, zero hunger, good health, clean water and energy, gender equality, sustainable cities, and quality education, etc.[1] (Sachs, 2015), there is an urgent need to mobilize the data revolution for *all* people in order to foster sustainable development and to hold governments accountable. More diverse, timely and trustworthy information can lead to better decision making and, in turn, it can enable individuals, public and private institutions, and companies to make choices that are good for them and for the world they live in. This is what the United Nations (UN) refers to as the *data revolution*, which has resulted in international standardized monitoring requirements being built into the UN's Sustainable Development Goals, as originally envisioned in a white paper "A World That Counts," commissioned by the UN Secretary-General (Independent Expert Advisory Group on a Data Revolution for Sustainable Development, 2014).

The Global Partnership on Sustainable Development Data[2] aims to respond to these needs: it consists of a global network of governments, non-governmental organizations, and businesses working together to strengthen and develop innovation efforts in a way that data is used to address the world's sustainable development goals. Specifically, the role of this network is to work to bring the resources of national governments, independent non-profits, and private companies to develop a framework of indicators to monitor the progress towards the achievement of the SDGs at the local, national, regional, and global level. A sound framework of indicators will turn the SDGs into a management tool to help countries and the global community in developing implementation strategies and allocating resources accordingly.

[1] Please see: http://www.un.org/sustainabledevelopment/sustainable-development-goals/.
[2] Please see: http://www.data4sdgs.org/.

However, a few important questions are left unanswered for policy makers, researchers, activists, and business people. First, how can we achieve transparency and accountability without endangering personal privacy or jeopardizing the business models of data-centric companies? In addition, how can we overcome the fact that a portion of the data required to measure poverty, diseases, human movements, and environmental conditions has been collected by and sits in the hands of private companies? Finally, what can we do to support a virtuous and sustainable data-driven process of policy making that involves people and ensures that their values and needs are met?

6.2 A New Social Contract

The answer we are developing to address the questions above is called OPAL (Open Algorithms).[3] The key idea is that rather than sharing private data with interested and authorized parties—which poses important technical, privacy, regulatory and social challenges (de Montjoye et al., 2018)—such parties can run approved algorithms to extract census-like data from behind private firewalls, such that privacy is preserved and the security and access control of the data are maximized (Hardjono et al., 2019). The obtained answers provide the knowledge required to better plan and monitor both government and private actions that affect individuals, communities, neighborhoods, cities, and entire countries.

We are carrying out country-scale experiments of OPAL in Senegal and Colombia with the help of Orange S.A., the World Bank, Paris21, the World Economic Forum (WEF), the Agence Française de Développement (AFD), and others. OPAL brings stakeholders together to determine what data—both private and public—should be made accessible and used, for which purpose and by whom. Doing so, OPAL adds a new dimension and instrument to the notion of a "social contract" (Gough, 1936), providing a forum and mechanism for deciding what levels of transparency and accountability are best for the society as a whole. It also provides a natural mechanism for developing evidence-based policies and a continuous monitoring of the various dimensions of societal wellbeing, thus offering the possibility of building a much deeper and more effective science of public policy.

In the rest of the chapter, we provide the readers with a compendium of the issues arising from the rise of data-driven approaches to policy making.

[3] Please see: http://www.opalproject.org.

Then, we turn to a more detailed description of the OPAL project, which can be seen as a step towards realizing the vision of a world where data and algorithms are used as lenses and levers in support of development.

6.3 The Rise of Data-driven Algorithmic Policy-making

Nowadays, many decisions with significant individual and societal implications previously made by human experts are performed or assisted by algorithms, in a variety of areas, such as lending, policing, sentencing, resource allocation, etc. (Willson, 2016).

Moreover, data-driven algorithmic decision making may enhance the overall efficiency of government processes and of public services delivery, by optimizing bureaucracy, providing real-time feedback and predicting outcomes (Sunstein, 2012). In his book *Technocracy in America* international relations expert Parag Khanna argues that a data-driven direct technocracy is a superior alternative to today's representative democracy, because it may dynamically capture the needs of the people while avoiding the distortions of elected representatives (Khanna, 2017). Indeed, human decision-making has often shown significant limitations in public policy, resulting in inefficient and/or unjust outcomes (Fiske, 1998). Thus, the adoption of data-driven algorithms could be seen as a demand for evidence-based decision-making and a better understanding of our individual and collective needs.

However, data-driven algorithmic policy making may generate some of the inefficiencies and harms highlighted in a book written by the development economist William Easterly (2014). In *The Tyranny of Experts* Easterly evaluates how poverty alleviation projects have often been governed by a "tyranny" of aid agencies, economists and think tanks who consistently favor top-down, technocratic governance approaches at the expense of the individual rights of citizens (2014). More specifically, Easterly underlines the cyclic nature of this tyranny: technocratic justifications for specific interventions are considered objective; intended beneficiaries are unaware of the black box decision-making involved; and experts act with impunity (2014). If we turn to the use of algorithmic data-driven approaches in the public sector we can draw several parallels with what we may refer to as the "tyranny of algorithms" (Lepri et al., 2017).

Researchers and activists have identified a range of social, ethical, and legal issues associated with algorithmic decision-making, including bias and discrimination (Pager & Shepherd, 2008; Barocas & Selbst, 2016), and lack of transparency and accountability (Pasquale, 2015; Zarsky, 2016).

For example, a study by ProPublica of the COMPAS Recidivism Algorithm, an algorithm used to inform criminal sentencing decisions by predicting recidivism, found that the algorithm was significantly more likely to label black defendants than white defendants, despite similar overall rates of prediction accuracy between the two groups (Angwin et al., 2016). In her book *Weapons of Math Destruction* the mathematician and activist Cathy O'Neil (2016) details several case studies on harms and risks associated with data-driven algorithmic decision-making, particularly in the areas of criminal justice and education.

Hence, we believe that it is not only important but also urgent to engage multi-disciplinary teams of researchers, practitioners and policy makers to propose, implement and evaluate in the real world algorithmic decision-making processes that are designed to maximize their fairness and transparency. In the next section, we describe one possible path that could be followed to address these challenges: the OPAL project (Hardjono et al., 2019).

6.4 The OPAL Project

The Open Algorithms (OPAL) project, developed originally as an open source project at MIT,[4] is now being deployed as a multi-partner socio-technological platform led by Data-Pop Alliance, Imperial College London, the MIT Media Lab, Orange S.A., and the World Economic Forum (WEF), which aims to open and leverage private sector data for public good purposes (Hardjono et al., 2019). OPAL projects have mostly been in two domains, the financial industries in Western countries (mostly MIT projects), and the leverage of mobile phone data for analysis and planning interventions in developing countries (mostly Data-Pop Alliance projects involving MIT, Imperial Orange, and WEF).

OPAL's first core feature is technological: the platform allows sending queries (i.e., running algorithms) to the partner companies' servers, behind their firewalls, so that raw data is never exposed. The second feature is socio-political: it involves co-designing the algorithms so that they are not only open but also respect local legal, regulatory and social standards as well as serve local needs. A key idea of OPAL is that algorithms should be verified by experts, policy makers, and citizens to be as free as possible from biases and unintended side effects. In order to obtain this, OPAL proposes to vet

[4] Please see: http://trust.mit.edu.

the algorithms running them on a specific data set and within a dedicated data repository. Once an algorithm has been vetted, it becomes a template that is digitally signed by the issuers (e.g., experts, public institutions, representatives of the communities that will be affected by algorithmic adoption and decisions, etc.). This template algorithm can be shared among a group of entities or even be published on a public site. However, we should highlight that this vetting process does not guarantee the quality of the output, which is instead a function of the quality of the input data.

The OPAL model of moving the algorithm to the data and of using vetting allows a data repository to choose whether it is willing to accept a submitted OPAL algorithm or not. If the data repository accepts a given vetted algorithm, it also has the option to impose additional filtering on the resulting data prior to being returned as answer to the issuer of the query (e.g., defining the degree of aggregation for a given answer). Moreover, each repository can introduce machine-learning algorithms as additional mechanisms to protect and respect privacy. Such algorithms allow a repository to detect if multiple accesses from the same entity may result in compromising Personal Identifiable Information (PII). Finally, blockchain technology can be used to capture and log both vetted queries and safe answers, thus providing a mechanism to support post-event audit and accountability.

The MIT version of the OPAL project is documented in the book "Trust: Data: A new framework for identity and data sharing" (Hardjono, Shrier, & Pentland, 2019) and it has been influential in many developed countries. For instance, the OPAL framework has guided development of EU government data governance, especially the architecture of Eurostat. In the United States (US), the OPAL framework formed the basis for a consortium of the eight largest US banks and Fidelity Investments to coordinate sharing of know-your-customer and anti-money laundering processes without sharing data. This consortium, called Akoya.com, had the result of both greatly increasing privacy and security as well as substantially reducing costs. In the first two years of Akoya.com almost 70% of US bank accounts and 50% of US credit cards became covered by Akoya's OPAL data framework.

The OPAL projects run by Data-Pop Alliance had the intent of helping to plan in developing countries and came out of the recognition that accessing data held by private companies (e.g., Call Detail Records (CDRs) collected by telecom operators, credit card transaction data collected by banks, etc.) for research and policy purposes has been a conundrum. To date for example, CDRs have been accessed and analyzed either internally, or externally through ad-hoc data challenges or through bilateral arrangements with a limited number of groups under Non-Disclosure Agreements.

These types of engagements have offered ample evidence of the promise, but they do not scale nor address some of the most critical challenges such as privacy, accountability, etc. (de Montjoye et al., 2018, 2014).

OPAL for safe use of CDRs has been deployed through pilots in Senegal and Colombia, where it has been endorsed by and benefits from the support of their National Statistical Offices and major local telecom operators. Local engagement and empowerment are central to the development of OPAL: needs, feedback and priorities have been collected and identified through local workshops and discussions, and their results will feed into the design of future algorithms. These algorithms will be fully open, therefore subject to public scrutiny and redress. A local advisory board is being set up to provide guidance and oversight to the project. In addition, training and dialogue will be organized around the project to foster its use and diffusion as well as local capacities and awareness more broadly. Initiatives such as OPAL have the potential to enable more human-centric accountable and transparent data-driven decision-making and governance, by involving a wide range of stakeholders in and through their design and implementation.

6.5 Conclusions

We live in an unprecedented historic moment where the availability of vast amounts of human behavioral data, combined with advances in machine learning, are enabling us to tackle complex problems through algorithmic decision-making. The opportunities to have positive social impact through fairer and more transparent decisions are paramount. In fact, this book, with its focus on big data in human development, reflects on such existing opportunities, but also describes some of the key challenges that we need to tackle before such opportunities can become a reality.

In this chapter, we have described a proposal that places humans and their societal values at the center of our discussions, as humans are ultimately both the actors and the subjects of the decisions made via algorithmic and/or human means. By involving people and ensuring that their values are met, we should be able to realize the positive potential of data-driven algorithmic decision-making while minimizing the risks and possible negative unintended consequences. We look forward to contributing to a world where—with the help of Big Data and related technologies—every person counts: their voices are heard, their needs are met, their potentials are realized, and their rights are respected.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review*, *104*, 671–732.

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., Rebaudet, S., & Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports*, *5*, 8923.

Blumenstock, J., Cadamuro, G., & On, R. (2016). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile phone data. Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 2014, 427–434.

de Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., de Cordes, N., Deletaille, S., Engø-Monsen, K., et al. (2018). On the privacy-conscientious use of mobile phone data. *Scientific Data*, *5*(1), Article 1. https://doi.org/10.1038/sdata.2018.286

de Montjoye, Y.-A., Kendall, J., & Kerry, C. F. (2014). *Enabling Humanitarian Use of Mobile Phone Data* (No. 26; Issues in Technology Innovation). Brookings, Center for Technology Innovation.

Easterly, W. (2014). *The Tyranny of Experts: Economists, Dictators, and the Forgotten Rights of the Poor*. Basic Books.

Fiske, S. T. (1998). Stereotyping, Prejudice, and Discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds), *Handbook of Social Psychology* (pp. 357–411). McGraw-Hill.

Gough, J. W. (1936). *The Social Contract*. Clarendon Press.

Gundogdu, D., Incel, O. D., Salah, A. A., & Lepri, B. (2016). Countrywide arrhythmia: Emergency event detection using mobile phone data. *EPJ Data Science*, *5*(1).

Hardjono, T., Shrier, D., & Pentland, A. (2019). *Trusted Data*. MIT Press.

Independent Expert Advisory Group on a Data Revolution for Sustainable Development. (2014). *A world that counts: Mobilising the data devolution for sustainable development*. http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794.

Khanna, P. (2017). *Technocracy in America: Rise of the Info-state*. CreateSpace Independent Publishing Platform.

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The tyranny of data? The bright and dark sides of data-driven decision-making for social good. In T. Cerquitelli, D. Quercia, & F. Pasquale (Eds), *Transparent Data Mining for Big and Small Data (Studies in Big Data)* (pp. 3–24). Springer.

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *PNAS*, *109*, 11576–11581.

Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Tatem, et al. (2016). Detecting climate adaptation with mobile network data in Bangladesh: Anomalies in communication, mobility and consumption patterns during cyclone Mahasen. *Climate Change*, *138*(3–4), 505–519.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit and consumer market. *Annual Review of Sociology*, *34*, 181–209.

Pasquale, F. (2015). *The Black Blox Society: The Secret Algorithms that Control Money and Information*. Harvard University Press.

Pastor-Escuerdo, D., Torres Fernandez, Y., Bauer, J. M., Wadhwa, A., Castro-Correa, C., Romanoff, L., Lee, J. G., et al. (2014). *Flooding through the lens of mobile phone activity*. IEEE Global Humanitarian Technology Conference (GHTC 2014).

Ruktanonchai, N. W., Bhavnani, D., Sorichetta, A., Bengtsoon, L., Carter, K. H., Cordoba, R. C., Le Menach, A., et al. (2016). Census-derived migration data as a tool for informing malaria elimination policy. *Malaria Journal*, *15*(273).

Sachs, J. (2015). *The Age of Sustainable Development*. Columbia University Press.

Smith-Clarke, C., Mashhadi, A., & Capra, L. (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2014)*, 511–520.

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of socioeconomic levels using cell phone records. *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization (UMAP 2011)*, 377–388.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, *14*(127).

Sunstein, C. (2012). *Regulation in an uncertain world*. National Academy of Sciences. https://www.whitehouse.gov/sites/default/les/omb/inforeg/speeches/regulation-in-an-uncertain-world-06202012.pdf

Traunmueller, M., Quattrone, G., & Capra, L. (2014). Mining mobile phone data to investigate urban crime theories at scale. *Proceedings of the International Conference on Social Informatics (SocInfo 2014)*, 396–411.

Wang, H., Kifer, D., Graif, C., & Li, Z. (2016). Crime rate inference with big data. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016)*, 635–644.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*(6104), 267–270.

Wesolowski, A., Metcalf, C. J. E., Eagle, N., Kombich, J., Grenfell, B. T., Bjørnstad, O. N., Lessler, J., et al. (2015). Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *PNAS*, *112*(35), 11114–11119.

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, Engø-Monsen, K., & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. *PNAS*, *112*(38), 11887–11892.

Willson, M. (2016). Algorithms (and the) everyday. *Information, Communication & Society*.

Wilson, R., Zu Erbach-schoenberg, E., Albert, M., Power, D., Gonzalez, M., Guthrie, S., Chamberlain, H., et al. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake. *Plos Current Disasters*, *8*.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, and Human Values*, *41*(1), 118–132.

7

Leveraging Big Data for Public Purposes in the Global South

LIRNEasia's Experiences

Sriganesh Lokanathan, Thavisha Perera-Gomez, and Shazna Zuhyle

7.1 Introduction

LIRNEasia is a think tank based in Sri Lanka but active across the emerging Asia-Pacific. Since 2012 it has been working to leverage big data for public purposes, representing one of the few institutes located in the Global South that conceptualizes and implements its own policy relevant big data analyses and research, bringing a Global South perspective to the global discourse on big data for development based on experiential evidence.

LIRNEasia's foray into the big data for development space started with exploratory research leveraging mobile network big data to inform public policy, particularly in relation to urban and transportation policy in Sri Lanka. While initially there was little to no demand for this type of work, LIRNEasia has built up expertise in this area over the years, carving out a niche and being recognized as a pioneer in the region (ITU, 2017). Using big data from a range of sources including mobile network big data, satellite imagery, CCTV footage, social media, coupled with traditional statistics from censuses and surveys, LIRNEasia's research now explores a wide variety of applications. These include, inter alia, analyses on human mobility and traffic, mapping land-use patterns, understanding communities (sub-national, national, and international), developing high frequency and high-resolution proxy indicators of economic activity and socioeconomic wellbeing of the population, modeling the spatial propagation of communicable diseases, etc.

LIRNEasia has also proactively sought to understand the implications of leveraging big data on privacy, competition and marginalization, as well as bias. This is especially useful to understand the utility of big data and

Sriganesh Lokanathan, et al., *Leveraging Big Data for Public Purposes in the Global South*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press.
© Sriganesh Lokanathan, et al. (2025). DOI: 10.1093/oso/9780198879145.003.0008

its application. LIRNEasia has also engaged in capacity building in local universities and among researchers to drive interest in big data for public purposes.

Yet the path has not been an easy one. LIRNEasia understands that such initiatives are challenging to replicate in developing countries. Data analytics is a frontier field where the tools and techniques are still being developed. Expertise, a critical input, is in short supply; the other being access to data. There are also challenges in making policy makers discerning consumers of such research. The challenges faced and lessons learnt are of relevance to others interested in developing their own policy relevant research practices in big data in the developing world, and perhaps also in the developed world.

7.2 LIRNEasia's Work

Some of the development problems that LIRNEasia has been interested in relate to urban issues: how can cities be made to be more livable? Can the spread of infectious diseases be reduced? Is there a role ICTs can play in this? How can innovative data and techniques be leveraged to answer these questions?

LIRNEasia conducts research using big data as well on the implications of the use of big data. The former deals with the application of big data techniques such as socioeconomic mapping using mobile network big data, while the latter addresses the socioeconomic implications of using big data, such as marginalization and bias in data and techniques.

7.2.1 Research Using Big Data

LIRNEasia's research leverages numerous big data sources including mobile network big data, satellite imagery, electricity data, social media data, CCTV footage, etc. In order to relate variables captured from these novel, high-frequency data sources to real phenomena of public interest, they need to be paired with more traditional data captured through censuses and surveys (see Figure 7.1).

For instance, LIRNEasia has sought to understand the degree to which mobile network big data could inform transportation planning for Sri Lanka's economic capital city of Colombo (Lokanathan et al., 2016), using it to understand the origin and destinations of the diurnal commuting population in the city. The results were closely aligned with existing transportation

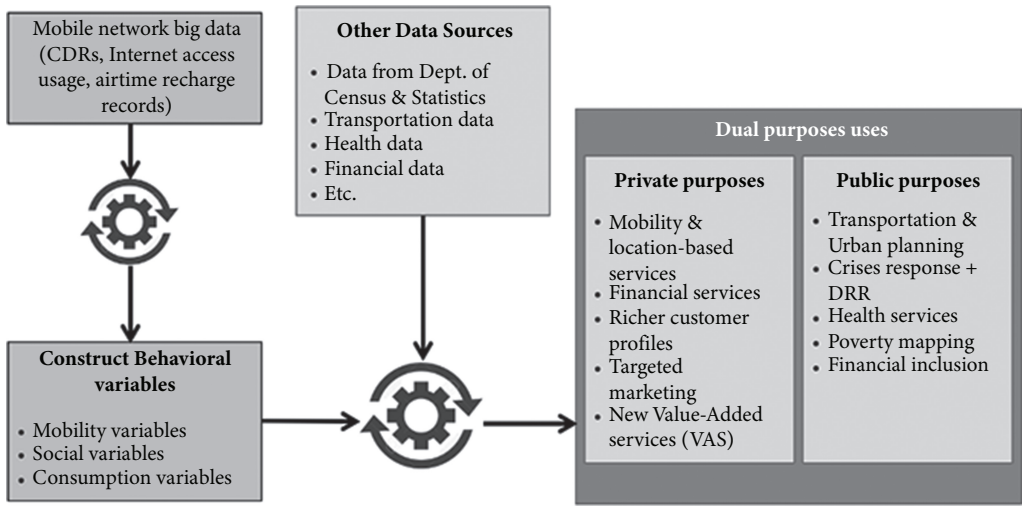


Figure 7.1 Combining mobile network big data with other data can help develop rich, timely insights that serve private as well as public purposes.

Courtesy of LIRNEasia analytics.

survey results. These and subsequent insights were used by the Western Regional Megapolis Project initiated by the Government of Sri Lanka in 2015 to develop the entire Western Province of the country into one contiguous economic region. The Western Province is Sri Lanka's largest province by population and accounts for at least 45% of the country's GDP. The insights developed by LIRNEasia were used to identify peripheral regions of the province for further development due to their high connectivity to the rest of the country. They were also used to better understand commuting and congestion patterns within the province. Once the plans were developed the newly created Ministry of Megapolis and Western Development continued to engage LIRNEasia to better understand mobility patterns for the local regions around the transit hubs that were being developed.

LIRNEasia also worked on leveraging mobile network big data to understand the impact of urban infrastructure. Preliminary work has shown that population-scale pseudonymized mobile network big data can provide new insights on the impact of urban infrastructure on travel behaviors in developing countries like Sri Lanka (Blumenstock et al., 2017). These can reveal detailed information about the short-term impacts of new transport infrastructure on urban congestion, and specifically route choices of travelers and the speed of travel across directly impacted roads. Given a longer panel of phone data, such data would allow for the quantification of the short- and long-term economic costs and benefits, as well as the distributional implications of these massive investments.

In 2018, LIRNEasia used call detail records (CDRs) to develop proxy indicators of poverty. These high-frequency indicators have utility in settings where official data are infrequent and/or difficult to obtain. This work used Sri Lanka as a case study to compare post-conflict areas with areas relatively untouched by conflict. The research showed that the post-conflict Northern Province exhibited different social and mobility patterns when compared to populations with similar socioeconomic characteristics in the Western Province, which was relatively untouched by the civil conflict of 1983–2009. For example in the Northern Province, those from a lower socioeconomic level were likely to speak more regularly to their contacts and travel further than those from a higher socioeconomic level. This is trend is reversed for those from the Western Province (Fernando et al., 2018).

Similarly, the team sought to understand the impact of human mobility when developing predictive models for the spread of dengue (Fernando et al., 2017). This was carried out by leveraging mobility variables derived from mobile network big data and incorporating additional data such as temperature, rainfall, and vegetation index, along with reported dengue

incidence data. The work, carried out in collaboration with the University of Moratuwa and the Epidemiology Unit of the Ministry of Health, showed that even in an environment like Sri Lanka where dengue is present and self-sustaining in most regions, the incorporation of mobility into the models improved overall predictions.

Further ongoing work (a collaboration with the University of Tokyo's Shibasaki & Sekimoto Lab) is articulating travel behavior and choice, including mode of travel, for the population disaggregated by economic and demographic attributes of travelers using novel mixed-method techniques that leverage mobile network big data, surveys, and mobile apps. The approach attempts to reduce bias and also account for different population groups amongst mobile phone users as well as non-users.[1]

Some of LIRNEasia's other work include understanding how communities (as inferred based on the geo-spatial communication and mobility patterns of the population) map with the actual administrative boundaries (Madhawa et al., 2015), utilizing country-to-country social media links (in this case Facebook friendship links) to provide proxies for trade and migration (Wijeratne et al., 2018); using electricity consumption data to understand optimum tariff brackets; automated analyses of CCTV footage to aid traffic management and traffic enforcement; and analyses of satellite imagery to understand land use characteristics, and also to predict the extant of crop cultivation and crop yield.

Two driving forces have influenced much of the initial analytics work undertaken at LIRNEasia. The first was to explore how data and especially alternate data sources could make cities in the Global South be more efficient and inclusive. The second was to develop a wide range of proof of concepts that could enlighten policy makers as well as the wider eco-system on the art of the possible.

7.2.2 Research on the Use of Big Data

There are many cases to be made for the use of high frequency data sources or transaction-generated data, commonly termed "big data," for commercial and public policy purposes. Given access and capacity (both in terms of intellect and hardware for the processing and analysis of such data sources) research can very well be carried out without understanding the "softer"

[1] A research paper is being developed, see: http://www.data4sdgs.org/news/dynamic-census-and-population-mapping-mobile-data-covid-19-response.

issues pertaining to its utility. However, as a research institute, LIRNEasia has sought to understand these nuances and the implications of using such non-traditional data, instead of focusing solely on churning results. As such, research to understand issues around privacy for example began as early as 2014. More recently the focus has been on algorithmic bias and ethics.

In addition to conducting analyses of big data, research has also been conducted on the applications and implications of big data research. As such, frontier issues of potential concern in the application of behavioral big data for public purposes, such as privacy (including group privacy), marginalization, adverse prioritization of services from improved poverty and wealth mapping (Samarajiva & Lokanathan, 2016) have been explored. By viewing these issues from a developing world context (i.e., a context of lower levels of datafication) this work has explored if and how the use of behavioral big data could aggravate these issues. In particular it is noted that marginalization has to be addressed as a special case of the problem of representativity. While representativity was neglected to some extent in the early days of big data, it is now receiving renewed attention by academics and practitioners.[2] Especially in the developing countries where datafication is incomplete, the question of "how well the population is represented," must be asked in relation to the questions being addressed. For example, when we leverage mobile network big data to develop proxy economic indicators of the socioeconomic standing of regions, we must be aware that while mobile phone ownership is high, both ownership and usage are biased towards those higher on the socioeconomic ladder. As such, being mindful of those who are "illegible" in the data such as the poor and the marginalized (who may not be able to afford a mobile phone) is essential so that analyses based on mobile network big data are sufficiently weighted to generalize their insights to the wider population rather than to only a certain segment of society.

LIRNEasia researchers have also studied bias from a technical perspective, specifically looking at how bias in algorithms and data affect real world decision-making. This research seeks to articulate algorithmic bias from a developing country context, taking into account the current state of limited datafication and also by examining ongoing instances where bias can manifest with the greater use of data in public policy decisions. The ultimate goal is to identify mechanisms (both technical as well as policy) to address issues of bias (Dias et al., 2020).

[2] For example, see Arora (2016).

7.3 Experience and Lessons Learnt

7.3.1 Influencing Policy

When LIRNEasia began conducting big data research in 2012, there was little to no demand for such work. While there are now increasing requests from various development initiatives within Sri Lanka, the journey from supply-push to demand-pull has been a long one. In the beginning, the team lead had to engage in a "policy enlightenment" process to create interest in big data research. LIRNEasia conducted several workshops at government institutions as part of its efforts to educate government officials on the uses and limitations of big data in their respective policy domains. At the same time LIRNEasia conducted workshops and public talks with various stakeholders including domain expert gatherings (e.g., the Young Planners Association), academia, civil society, as well as the public. Although there was no demand for insights when research began in 2012, early research enabled the team to capitalize on policy windows as they emerged. Each policy engagement built on previous relationships and primed the team for potential future opportunities, such as the contributing to the Western Region Megapolis Planning Project.

Because the big data work was new and novel, it was difficult to assess the quality of the research. It was here that the fact that LIRNEasia had an established prior presence (albeit in areas other than big data analytics) lent itself towards building credibility of the big data work the team was engaging in. However, while these initial successes in being able to influence policy have been encouraging there is more to be done if the responsible use of big data is to be mainstreamed in government. First, we need a much larger cadre of informed and discerning consumers of big data amongst government institutions. But it is one thing for government officials to be enlightened on the art of the possible in relation to big data, but another to make the responsible use of big data in their workstreams the norm. As such, the path to mainstreaming will require collaborative pilots so that there are elements of "learning by doing" which enables them to internalize *precisely* how big data can aid and complement their work.

7.3.2 Data

Because LIRNEasia does not generate any data of its own, it depends on other sources to obtain relevant big data for its analyses. When the work began in 2012 there were no established processes or data partnerships. However,

LIRNEasia's initial work in ICT policy and regulation created the credibility to discuss data access with the senior leadership of mobile network operators. Yet the process to actually obtaining access to pseudonymized data was a long and laborious one. The negotiations began with CEOs of MNOs to obtain in-principle agreements and then long-drawn-out negotiations around the specifics with the second and third tier management over the course of many months (sometimes over six months) before an agreement was signed. At each point, mutual benefits such as the potential for MNOs to adapt our methods for deriving insights for their commercial purposes were emphasized. Two of the concerns that were proactively dealt with were privacy and the risk of disclosure of personal information, and potential breach of existing laws and regulations.

To address privacy concerns, it was suggested that the MNOs pseudonymize the data (i.e., remove all personally identifiable information), prior to providing access. Non-disclosure agreements were also signed with each MNO, which gave the operators the option of signing off on the results of the analyses before public release. Furthermore, all researchers working with the data signed individual non-disclosure agreements with LIRNEasia for the duration of the research engagement. Once the pseudonymized data was obtained and mounted on our own in-house cluster, access to the data was also restricted through physical security guidelines.

While accessing data from the private sector posed significant challenges, accessing government data was at least as challenging. Government agencies were more at ease sharing data directly with other government institutes rather than think tanks. Hence collaborations with universities, which in Sri Lanka are essentially government entities, helped us to address data access challenges. For example in our work on understanding how mobile network big data could be used to improve the modeling of the spatial spread of dengue, a tri-party collaboration was formed between LIRNEasia, the University of Moratuwa and the Epidemiology Unit of Sri Lanka (under the Ministry of Health). This enabled the fusion of data from different sources, thereby enriching the analysis.

7.3.3 Skills

One of the principal challenges faced by LIRNEasia has been human resources. Data scientists are in short supply in Sri Lanka, and attracting and retaining staff with strong computer science skills remains difficult, particularly given the fact that those coming out of Sri Lankan universities are

often attracted to higher-paying private sector jobs. LIRNEasia has been able to overcome these challenges by building teams with complementary skills: computer science and data mining, statistics, and domain knowledge. These multi-disciplinary teams have helped us overcome the short supply of inter-disciplinary data scientists. Analytical thinking is prioritized over knowledge of big data tools in our recruitment interviews.

One of the key selling points for us is that working at a think tank specializing in big data research provides a unique stepping-stone for admission into good PhD programs. To that end, researchers have opportunities for publication and conference papers, and encourage individual researchers to build their "brand." While this approach has helped us attract talent with a deep interest in the big data space, there is the inevitable churn as data scientists leave for their PhD programs.

The loss of continuity as researchers move onto higher studies has an adverse impact on knowledge transfer. One way to minimize this has been a specific emphasis on thorough documentation of methods, results, challenges, etc. As such, researchers engage in documenting their analysis, to the extent possible, to ensure that other researchers can pick up where they left off. Moreover, those who leave for higher studies make themselves available to new researchers, acting as advisors on projects.

During the course of research the need for specialized expertise was evident. To that end, in September 2015 LIRNEasia constituted an advisory group for its ongoing big data for development research, consisting of international and local experts from a variety of domains. Given the multi-disciplinary nature of LIRNEasia's big data for development research, the advisory council was intentionally constituted from different disciplines and domains. Members included data science experts, sector specific experts (e.g., transport planning, urban planning), social science experts, and legal experts. The purpose of the advisory group was multifold. Council members helped hone our research agenda in addition to bringing in new research ideas. They served as important reviewers for our ongoing work, whilst also collaborating on research with our team, which aided our own capacity development. Additionally, they strengthened our networks, facilitated collaborations and helped identify funding sources. Research collaborations with members of LIRNEasia's advisory group, as well as internships with carefully selected PhD students from institutions such as MIT and University of Tokyo has led to mutually beneficial partnerships, enabling collaborators to further their own research agendas with access to novel datasets and linkages to the policy environment, whilst helping LIRNEasia to further develop the skills of our team.

7.3.4 Strategic Engagement and Catalysis

Ensuring that the team engages in local, regional, and global research has helped to influence such public, practitioner, and academic discourses with experiential evidence. At the same time, these conversations have helped to calibrate the research to ensure that it remains relevant. LIRNEasia's CEO was appointed to the board of the Global Partnership for Sustainable Development Data (GPSDD) and the co-founder and former team leader of our big data research was a founding board member of Citra Social Innovation Lab, Sri Lanka's first social innovation lab, as well as a Research Advisor to Pulse Lab Jakarta. Moreover, LIRNEasia's Founder Chair was appointed to a second term on the Privacy Advisory Group of the UN Global Pulse.

LIRNEasia has also been engaging with academics, practitioners, and policy makers in the region, sharing our thoughts on big data for development and governance. The focus has been on highlighting the need for effective, fit-for-purpose public policy and regulation for big data, including algorithms. Over the last several years, the framing of the development discourse has shifted steadily to the Sustainable Development Goals (SDGs), and the team has been active in these discussions. LIRNEasia conducted a scoping study on identifying the specific uses of big data in the SDGs, outlining non-traditional data sources that have been leveraged as well as methods utilized to produce insights of relevance to the global goals and the associated targets (Lokanathan et al., 2017). In March 2018, LIRNEasia was proud to partner on Sri Lanka's first National Symposium on Data for Sustainable Development. The event offered a unique national platform to share our views and experiences on the use of data, and big data in particular, for the SDGs. More recently, a session at the World Data Forum on leveraging South–South Cooperation to harness the power of new data sources was organized.

LIRNEasia has also helped initiate the International Development Research Center's (IDRC) Big Data for Development Network, established in early 2018, and continues to be a part of it. The network currently includes five nodal institutions covering Africa, Asia, and Latin America, and is meant to foster big data for development work in the Global South and to bring more Southern voices into the international discourse.

Whilst upgrading our own capacity on a continuous basis, LIRNEasia continues to help develop local capacity to conduct big data analytics and conduct data-driven decision making for social good. The team has mentored and supervised undergraduate and masters students from multiple

universities in Sri Lanka, while also engaging with academics on research. LIRNEasia helped found the establishment of a center of excellence at the University of Moratuwa, Sri Lanka's premier engineering university, to conduct big data analysis for both the private as well as public sectors. The center, DataSEARCH ("Data Science, Engineering, and Analytics Research Hub") is a multi-disciplinary research center, and some of its initial research extends LIRNEasia prior collaborations with different domain experts at the University. LIRNEasia was also a Founder Member of Citra lab, Sri Lanka's first social innovation lab which is a joint initiative between the Ministry of Science, Technology and Research, and the United Nations Development Programme. LIRNEasia continues to actively engage and foster the development of these two initiatives. The former Team Leader of Big Data Research was a founding member of DataSEARCH and also sat on the board of Citra. The team also continues to mentor and advise social scientists in Sri Lanka on leveraging big data through an initiative spearheaded by the National Science Foundation (NSF).

Through ongoing collaborations with Pulse Lab Jakarta and UNDG for Asia Pacific, LIRNEasia is also engaged in advising UN data projects in the Asia Pacific region as well as conducting sessions at UN-sponsored training workshops for UN staff in the Asia Pacific region.

7.4 Way Forward

Building on initial successes in informing policy and engaging with the global discourses, the next steps for LIRNEasia are clear—namely, to mainstream the responsible use of big data within Sri Lanka and catalyze such activities across the Global South. Mainstreaming such work in a sustainable fashion will require the confluence of a variety of actors, and as such the strategy will be to help develop the broader ecosystem by helping to support new and emerging institutions and also through broader collaborations with academia and government. Similarly with the global discourse shifting from big data narrowly to more broadly algorithms and Artificial Intelligence, LIRNEasia will again build on its prior work and work with partners in the big data for development network, to bring experiential evidence with a Global South perspective into the global discourse.

References

Arora, P. (2016). Bottom of the data pyramid: Big data and the global South. *International Journal of Communication*, *10*, 1681–1699.

Blumenstock, J., Maldeniya, D., & Lokanathan, S. (2017). Understanding the impact of urban infrastructure: New insights from population-scale data. *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, 1–12. https://doi.org/10.1145/3136560.3136575

Dias, V., Lokanathan, S., & Wijeratne, Y. (2020). *A Brief Primer on Bias in Machine Learning and Algorithmic Decisions*. LIRNEasia. https://lirneasia.net/wp-content/uploads/2020/05/A-Brief-Primer-on-Bias-in-Machine-Learning-and-Algorithmic-Decisions.pdf

Fernando, L., Perera, A., Lokanathan, S., Ghouse, A., Tissera, H., & Samarajiva, R. (2017). *Impact of Human Mobility on Spread of Dengue in Sri Lanka*. http://netmob.org/assets/img/bookofabstract_oral_2017.pdf

Fernando, L., Surendra, A., Lokanathan, S., & Gomez, T. (2018). Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka. *Proceedings of the Fourth International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets*, 1–12. https://doi.org/10.1145/3220547.3220549

ITU. (2017). *Measuring the Information Society Report 2017*. ITU. https://www.itu.int:443/en/ITU-D/Statistics/Pages/publications/mis2017.aspx

Lokanathan, S., Kreindler, G. E., Silva, N. H. N. de, Miyauchi, Y., Dhananjaya, D., & Samarajiva, R. (2016). The potential of mobile network big data as a tool in colombo's transportation and urban planning. *Information Technologies & International Development*, *12*(2), 2. https://itidjournal.org/index.php/itid/article/view/1506.html

Lokanathan, S., Perera-Gomez, T., & Zuhyle, S. (2017). *Mapping Big Data Solutions for the Sustainable Development Goals*. LIRNEasia.

Madhawa, K., Lokanathan, S., Maldeniya, D., & Samarajiva, R. (2015). *Using Mobile Network Big Data for Land Use Classification*. Communication Policy Research South Conference.

Samarajiva, R., & Lokanathan, S. (2016). *Using Behavioral Big Data for Public Purposes: Exploring Frontier Issues of an Emerging Policy Arena*. LIRNEasia & Open Society Foundation.

Wijeratne, Y., Lokanathan, S., & Samarajiva, R. (2018). *Countries of a Feather: Analyzing Homophily and Connectivity Between Nations Through Facebook Data. TPRC*. TPRC 46: The 46th Research Conference on Communication, Information and Internet Policy 2018.

8

Advancing Human Development with Artificial Intelligence

Experimental Work from Uganda

Paula Hidalgo-Sanchis

8.1 Big Data and Human Development in Uganda

The indicators of Human Development have progressed slowly in the last 30 years in Uganda. The country's Human Development Index (HDI) for 2018 was 0.528, positioning it at 159 out of 189 countries and territories. However, taking into consideration inequality in the three dimensions of the HDI, the HDI falls to 0.387 (United Nations Development Programme, 2018).

Between 1990 and 2018, some dimensions of the HDI improved, such as an increase by 17.1 years of Uganda's life expectancy at birth (from 45.9 to 63.0); an increase in the mean years of schooling by 3.3 years (from 2.8 to 6.1); and an increase in expected years of schooling by 5.6 years, from 5.7 to 63.0 (United Nations Development Programme, 2018). However, as inequality in a country increases, the loss in human development also increases.

The latest human development report for Uganda (United Nations Development Programme, 2016) highlights that despite progress at the national level, large parts of society have missed out, and people's opportunities are still dependent on their place of birth. The report analyses how people born in the Northern region have less access to fair choices to enhance their capabilities due to the consequences of the recent civil conflict in the region.

To address development challenges, Uganda was one of the first countries to approve a National Development Plan (Government of Uganda, 2015) in line with the UN's Sustainable Development Goals (SDGs). Sixty-nine percent of the SDG targets are reflected in the plan.

To achieve the 2030 Agenda for Sustainable Development and leave no one behind, all voices need to be heard, and bridges need to be built to bring

Paula Hidalgo-Sanchis, *Advancing Human Development with Artificial Intelligence*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press.
© Paula Hidalgo-Sanchis (2025). DOI: 10.1093/oso/9780198879145.003.0009

people's voices into the center of policies. We live in a period of human history in which digital data is generated without interruption everywhere, including in Uganda. New technologies are now available to tap into the digital footprints generated as people go on with their daily lives. To use these technologies and tap into new data sources, researchers need to look at data with a new mindset. With AI technology, unprecedented large volumes of people's voices are now accessible in real time to inform policies and improve people's lives—allowing government and development partners to include the voices of vulnerable population groups, people living in remote areas, and groups excluded from societies.

While there is a digital divide in countries like Uganda where only part of the population is online, the digital divide is vanishing. There is still inequality in accessing digital devises linked to gender, location, age, literacy rates, and socioeconomic status, but access to digital devices in the country is growing faster than access to clean water or to decent jobs.

Indeed, access to mobile phones is growing exponentially, and with it, access to the Internet. In Uganda, there has been an explosion in the use of mobile phones since the first network went live in 1995. In 1999, the country became the first in Africa where the number of mobile subscribers passed the number of fixed-line users, and the ratio was more than 18:1 in 2017 (Internet World Stats, 2020). Internet access has a penetration rate of 37.9 of the population nowadays (Uganda Communications Commission (UCC), 2018b). Mobile phones are driving a digital revolution in Uganda with 71% of Ugandans owning a mobile phone (National Information Technology Authority (NITA) Uganda, 2018) and 25.6 million mobile subscriptions in 2018 (Uganda Communications Commission (UCC), 2018b).

The opportunities and challenges of applying big data and Artificial Intelligence (AI) technology to promote human development are certainly unfolding, as the digital trail of every aspect of human life can now be counted, measured, and analyzed. A young team of data scientists, data engineers, social scientists and artificial intelligence experts, working at Pulse Lab Kampala, a Lab of the United Nations Global Pulse network, has experimented with technology and new methods to tap into new datasets and inform human development. Between 2016–2018, the team experimented to test if the methods used by the private sector to gain market intelligence from public discussions over digital platforms could be used to advance human development.

To do so, the team at Pulse Lab Kampala and partners developed a technology prototype based on AI to analyze public discussions in the main media platform in Uganda, which is public radio. The results of the experimental

work inspired researchers around the world about the possibilities of big data and AI applied to human development and also sparked a debate about ethics on AI (Domonoske, 2016; Hidalgo-Sanchis, 2021).

8.2 Analyzing Public Radio Content with AI Technology

8.2.1 Why Radio?

According to UNESCO (2014), radio is the most reliable and affordable medium of accessing and sharing information in Africa. In Uganda, as in many other countries in East Africa, radio is used by citizens as the main channel to voice opinions and share reports about incidents that affect their lives and communities (Madamombe, 2015). Community radio stations are used similarly to social media in Uganda. There are 292 operational FM radio stations in the country (Uganda Communications Commission (UCC), 2018a) and approximately 60% of radio programs are conducted with participation from listeners who call in to express their views and opinions. Most radio stations host talk shows or studio discussions involving local officials and community leaders, which are popular among all ages and especially in rural areas. Pulse Lab Kampala mapped out the content of radio programs for 92 radio stations during 2017 and, based on this, estimated that 20,000–25,000 people participate in radio shows in Uganda every day.

The anonymized millions of words that are spoken every day on public radio are a source of big data that can be used to improve people's lives and advance human development. Public radio discourse is of high qualitative value to inform social policies because it is anonymous and unconstrained. While all data is biased by different factors, the empiric research conducted over the years, revealed that public radio discourse includes testimonials and personal views expressed without fear of judgement from a researcher that are very valuable to inform programs and policies (World Health Organization, 2021). For example, an anonymous citizen expressed this in relation to gender equality and gender violence:

> A woman is the one who must show respect to a man . . . it's me who marries the woman from her parents' home and she comes and to my home primary to respect me . . . I stand on my two legs and say that women must respect men. The women are from the ribs of men. Second, the man is head of the woman for this reason, the woman must respect the man and if she does not do that, it brings violence within the family. [Radio talk show from the Northern region of Uganda, 2018]

Public discourse on radio is also valuable to advance human development because citizens share first-hand reports of incidents that are not recorded elsewhere. For example, in Uganda, losses associated to natural disasters are common, but they are under-reported at local, regional, and national levels, and this hampers the design and implementation of policies:

> Lightning struck 20 cows to death in Guda Palwo village in Lamwo district. Owners say that they had just returned from grazing and it has cost them up to 22 million shillings and they had got them from the project. [Radio talk show from the Northern region of Uganda, 2016]

8.2.2 Frontier Technology

With the aim to analyze public radio content in Uganda, Pulse Lab Kampala developed a prototype based on AI using machine learning, neural networks, and deep learning methods (see Figure 8.1). The prototype was considered a worldwide innovation in the field of international assistance, and was built using similar techniques to those developed by companies such as Google or Facebook for complex face recognition methods (Hidalgo-Sanchis, 2016; Kirkpatrick, 2021; World Health Organization, 2021; Hidalgo-Sanchis, 2021). Many young Ugandans with no previous experience developing these technologies participated in the experimental work, demonstrating the wealth of human capital in the country.

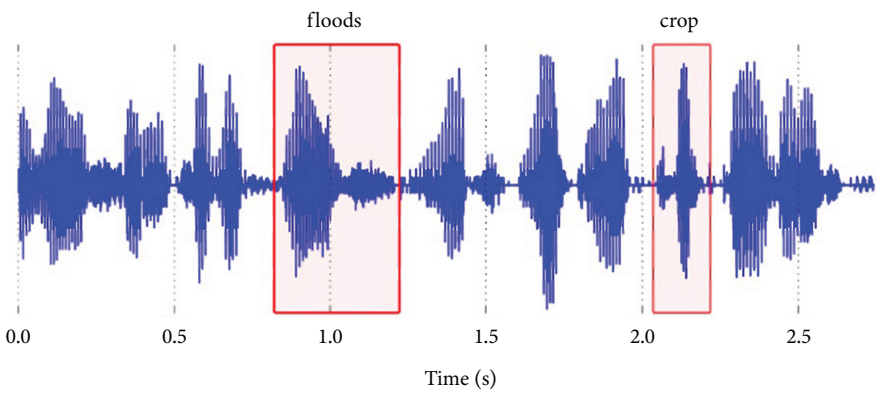


Figure 8.1 Transcription from audio into text by Pulse Lab Kampala.

Courtesy of Pulse Lab Kampala—part of the UN Global Pulse Network.

Note: The lines represent speech from public radio, the box represents the extract of audio and the words above are the transcription of the audio into text (translated into English)

Through the in-house technology, public radio content is streamed via IT equipment deployed in the field in a processing pipeline with a secure API. The radio stream is then ingested in cloud servers and goes through a series of filters that allow selection of content. As around half of radio station broadcasts consist of music, a first filter is applied to take out music from the audio stream. The radio stream then goes through a speech-to-text module built based on a large-vocabulary Automatic Speech Recognition (ASR) system (Agenbag & Niesler, 2019) that uses multilingual acoustic modelling (Menon et al., 2017). The module provides a full transcription of the audio content within the confines of a defined vocabulary. This module is based on Natural Language Processing (NLP) and was developed with partners from the Department of Electrical and Electric Engineering at Stellembosh University of South Africa. The ASR transcribes audio in Luganda and Acholi, two vernacular languages of Uganda, into text.

The ASR outputs, consisting of text transcripts (including every word detected and precise start and end times) are streamed into the next module. Analysts can then extract content with data-mining techniques using queries defined with selected keywords and key phrases (Menon et al., 2018), for example, floods, refugees, or health. The extracted data is later assessed by analysts who select relevant content.

As a result of the process, which is conducted in near real time (less than 24 hours from air to dashboard), a compilation of anonymous testimonies, options, and community reports related to the topic of interest are available for analysis. An overview of this process is depicted in Figure 8.2.

8.3 Testing the Technology

8.3.1 Methodology

The team from UN Global Pulse conducted applied research with partners over three years (from 2016 to 2018) to assess the value of the insights extracted from public radio data to advance different dimensions of human development Uganda (UN Global Pulse, 2017). The applied research consisted of a series of case studies conducted along with partners experimenting with different methods of analysis, such as categorizing radio content by topic, by type of speaker, or analysing trending topics over time. The case studies focused on different topics, including the analysis of social cohesion during an influx of refugees into Uganda from South Sudan (UN Global Pulse, 2018), HIV/AIDS, infectious diseases, quality of education, and natural disasters.

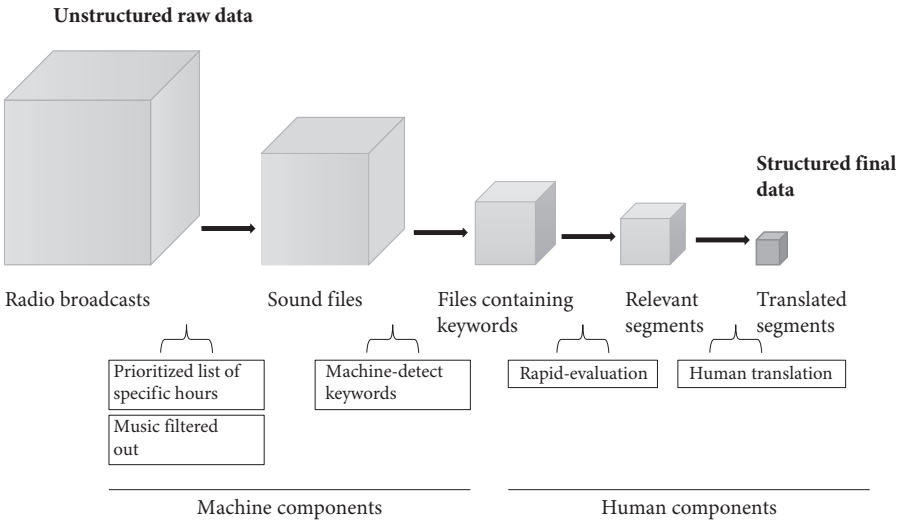


Figure 8.2 Converting unstructured audio data into structured data by Pulse Lab Kampala.

Courtesy of Pulse Lab Kampala—part of the UN Global Pulse Network.

The applied research complied with the Article 27 of Ugandan's Constitution *Right to Privacy of Person, Home and Other Property*, the Data Protection and Privacy Act, 2019 (the 2015 Data Privacy and Protection Bill at the time when the research was conducted), and the policies and regulations in the communications sector defined by the Uganda Communications Commission. To ensure the protection of digital rights and the ethical use of technology, the team's work was guided by the Data Privacy and Protection Principles formulated by the Data Privacy Advisory Group lead by UN Global Pulse and the Risks, harms and benefits assessment tool produced by the group. Informed by these, the following protocols were established: the technical architecture and components of the technology were only accessible to staff abiding by the contractual rules and regulations of the United Nations; and the analysis results were anonymized and presented at a level of aggregation that did not allow the re-identification of citizens.

The case studies were conducted with a combination of automated software processes and human analysis stages. In the first stage of human analysis, analysts categorized manually the radio content of 92 radio stations with the objective of targeting potentially valuable radio content and to reduce the volume of data to be processed by the prototype. The team developed a "gold list" that included talk shows with grassroots discussions and people phoning in the studio and a "silver list" that included radio programs with studio discussions among community leaders, government officials and the public in the

radio studio. Based on these categories, the system was set up to filtered out automatically content in the "bronze list," which included programs of news reports and in the "black list," which included programs on sports, celebrity gossip, and music.

In the next automated software stage, the streamed content passed through the ASR filter and through filters to mine the text with queries including required, optional, and to exclude keywords and key phrases. In the next stage of human analysis, the content of the ASR output was assessed and categorized by researchers combining quantitative (analysis of volume of data per topic and per type of listener over time) and qualitative (detecting rumors and testimonials of events at local level) analysis methods. Biases were corrected in the data on topic via: triangulation of reports from individuals with those of other listeners; aggregation of analysis results to correct fake reports, testimonials or opinions; introduction of a margin of error to correct false data or human biases; and interpretation of analysis results, considering that radio content is commonly sponsored and thus might be subject to self-censorship.

Through this process a large and unstructured dataset, containing largely irrelevant discussion, was transformed into a smaller, structured dataset containing categorized text of only the topics of relevance. In the pilot study conducted from October 5, 2016 to January 25, 2017, which focused on the quality of public service in the Central and Northern Regions of Uganda, 16 hours of relevant radio content were extracted from over 100,000 hours as depicted in Figure 8.3.

8.3.2 Results

The results of the applied research showed that public radio content analysis can support the following critical areas to advance human development in the country, namely: quality of public service delivery; societal behavioral change; and public sector accountability. A high volume of rich public discussions and citizen's testimonials were continuous on air on these areas, with high potential to support the design and monitoring of policies.

8.3.2.1 Quality of Public Service Delivery

Quality of public service delivery remains a challenge in Uganda specially in the health sector, where absenteeism of medical staff, poor facilities and medical equipment, shortage of medicine stocks and neglecting patients are common across the country. Improved service delivery frameworks in the

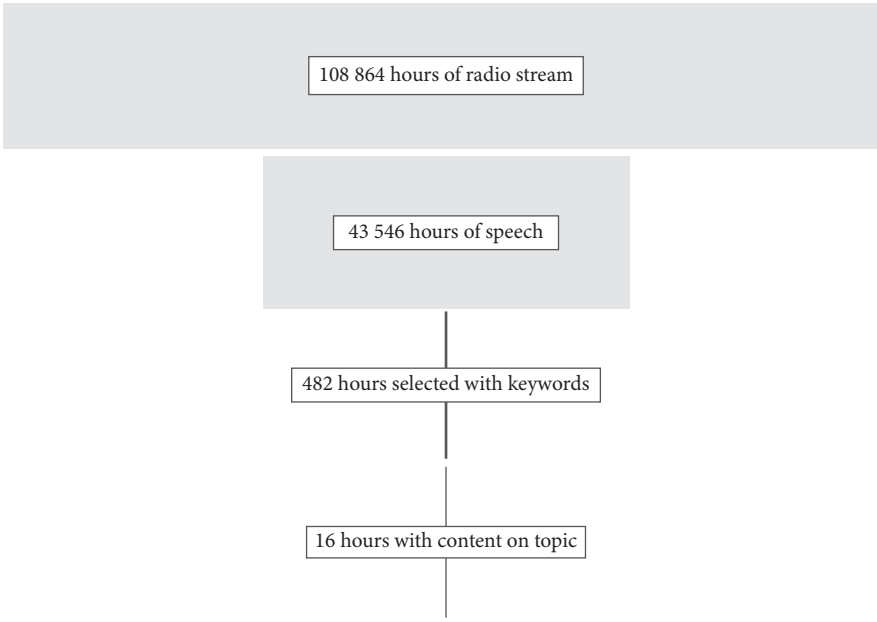


Figure 8.3 Filtering process of public radio content with AI.
Courtesy of Pulse Lab Kampala—part of the UN Global Pulse Network

country require stronger links between the policy makers who define targets, those who deliver the service, and the citizens who receive them (World Bank, 2009).

Reports, testimonials, and discussions about the poor quality of health public service delivery were a trending topic on radio in Uganda during the research period. The team at Pulse Lab Kampala analyzed radio streams along with the Ministry of Health of Uganda and concluded that the unsolicited public discourse available on radio is a measure of client satisfaction on the quality of public sector delivery. These are some examples of public discourse on the topic:

> If you also go to the health center in Omugo, there are no drugs when you go there for treatment" . . . "We have a problem in Ongango sub county, you find that if they have brought medicine today, the following day . . . the medicine is over" . . . "Our problem with the hospital is that there are no drugs in them and in most cases, you only get Coartem and Panadol yet those are not the only diseases that we suffer from." [Northern Region, 2016]

> The people that they put to work in the health center are nowhere to be seen, the doctor for that health center is not available. [Central Region, 2017]

> She delivered . . . a baby girl . . . when they brought that baby back . . . the baby is a boy . . . how comes that you people bring a dead boy? I need my daughter. Then from there shortly that doctor explained to her that now if you want me to go and look for that baby, give me twenty thousand (Ugandan Shillings), so that I am in position to go and look for your baby. [Western Region, 2018].

8.3.2.2 Societal Behavioral Change

Advances in human development in Uganda face challenges due to societal behavior. Myths, inherent beliefs, spread misconceptions and rumors play an important role hindering the effectiveness of policies on areas as HIV/AIDS, gender-based violence (GBV), disease prevention, reproductive health, and malaria.

Some of the challenges in these areas include: (1) there are 1.2 million people (aged 16–65) living with AIDS in Uganda and the country had rising prevalence of HIV rates in the last years (Government of Uganda, 2017); (2) despite the conducive legal and policy environment GBV is still high and there are alarmingly high rates of domestic violence affecting half of women (United Nations Women, 2020); (3) neglected diseases that can be eradicated by adopting new behaviors, such as blinding trachoma and schistosomiasis, are still spread in Uganda (Uniting to Combat NTDs, 2018); (4) the implementation of sexual and reproductive health policies in Uganda face challenges affecting adolescent girls, critically; and (5) malaria is the main cause of morbidity and mortality in Uganda (Ministry of Health of Uganda, 2020).

To address these challenges, behavioral change campaigns have been implemented in the country for decades with an important component on radio. In 2015 for example, the Uganda Health Marketing Group (2020) and the United Nations Population Fund (UNFPA) ran the social and behavior change communication campaign *If it is not on, it is not safe* promoting the use of condoms among youth (between 18–24 years). In 2018, the Ministry of Health and partners in civil society launched the campaign *Silence is Deadly* to encourage young people to speak out on sexual and reproductive health issues affecting them.

The research conducted by Pulse Lab Kampala and partners revealed that the analysis of myths, inherent beliefs, misconceptions, and rumors exposed on public radio discussions can support monitoring of the impact of societal behavior campaigns. These are some examples of public discourse on the topic:

> Even if most people think that immunization is for young children, even old people can be immunized. When can all be immunized? [Central region 2016]

> This kind of medicine that is being sprayed, that medicine for mosquitoes . . . I don't think is good for us. Because when they do not spray the medicine, there are no mosquitoes, but once they spray, there are so many mosquitoes. We need a clear explanation, because all these things are going to kill us the Acholi people and in the near future, we shall be finished. [Northern region, 2017]

> These medicines as per what they say are the ones sometimes making the pregnant women die . . . I am very sad with these artificial family planning [methods] administered that you see here. These things I see they contribute greatly; they kill our young women as well as the mature ones. [Western region, 2018].

8.3.2.3 Accountability

Uganda is the largest refugee-hosting country in Africa, with an estimated 1.4 million refugees (United Nations Human Rights Council, 2020) from South Sudan, the Democratic Republic of Congo, Burundi, Somalia, Rwanda, Eritrea, Ethiopia, and Sudan. In February 2018, the national and international media exposed allegations related to corruption on the refugee response, including inflated (refugee) numbers, double counting of refugees, extortion, and child trafficking (e.g., South Sudanese girls being sold as wives to combatants with the possible complicity of Ugandan authorities). Successive investigations in the country have probed these allegations.

The research by Pulse Lab Kampala and partners found evidence of public discussions on corruption related to refugees over three months before the news hit the international media. The team concluded that the aggregated analysis of anonymous testimonials aired on public radio can feed accountability and risk management systems. Some examples of public discourse on the topic include:

> Some of the refugees are tricking and write their name two times. [Western region, 2017]

> There were donations to help with buying seedlings for planting to assist the refugees with food . . . only a few people were given these seedlings and the rest of the seeds were used in personal farms of these government officials. [Northern region, 2017]

> Two of the youths gave in their statement, confessing that truly they escaped from that training ground . . . the joint security forces arrested him for further investigation and prosecution. During the interrogating [X] admits that truly he has been recruiting the youth to go and participate in a training in South Sudan to join the rebel forces. [Western region, 2017]

8.4 Conclusion

Innovative research methods are necessary to advance all dimensions of human development, including human security, equity, participation, and freedom. The ongoing digital transformation of societies offers unprecedented opportunities to tap into new data sources to inform the design of policies and monitor their implementation and their impact.

AI-based technology can transform a large stream of unstructured public radio discourse into a reduced, structured, rich dataset. This dataset is a source of primary data that surpasses any volume of primary data collected by any traditional method to gather people's voices (for example, surveys), it is available for analysis in near real time (less that 24 hours from air to dashboard), and it contains citizen's testimonials and local reports not biased by questions asked by researchers.

The applied research showed the value of public radio content analysis to support the following critical areas to advance human development in the country: quality of public service delivery; societal behavioral change; and public sector accountability. The near real time analysis of radio streams contains citizen's testimonials and local reports that can support the design of monitoring and impact evaluation of policies in these areas. After the research, Pulse Lab Kampala continued supporting the UN system and the Ugandan government in analyzing public radio content to different dimensions of human development, for example in supporting the Spotlight initiative to eliminate violence against women and girls (UN Global Pulse, 2019). The technology has been adopted for other uses outside Uganda, such as infodemic monitoring in Africa (Jun, 2022) and for efforts to support peace building in Mali (Lemm, 2021).

Analyzing public radio data presented challenges associated with the limitations and biases of the data. While the dataset is rich, its representativeness is limited to those who participate in radio shows. Addressing the biases of the dataset was complex and required in-depth knowledge of public radio content in Uganda and the interpretation of analysis results in light of its inherent biases.

The team experienced challenges in the development of pronunciation dictionaries and acoustic models to build the ASR for Luganda and Acholi, as they are low-resource languages. Creativity, imagination, and teamwork are essential to overcome them. The scale of the datasets analyzed challenged the researchers to design methods of analysis to make sense of large volumes of qualitative data, and this resulted in the development of software tools to facilitate the process.

The technology and methods used to analyze public radio data can be used in other countries across Africa, and other continents where public radio is used in a similar way to Uganda. To do so, disciplinary teams of social researchers and technologists are required.

Experimental work such as that led by Pulse Lab Kampala reveals new challenges in the area of data privacy and protection and ethical use of technologies. Scaling up the use of the technology presented requires the implementation of protocols for analysis and the presentation of the analysis results informed by ethical frameworks.

References

Agenbag, W., & Niesler, T. (2019). Automatic sub-word unit discovery and pronunciation lexicon induction for ASR with application to under-resourced languages. *Computer Speech & Language*, *57*, 20–40. https://doi.org/10.1016/j.csl.2019.02.002

Domonoske, C. (2016, October 31). Turn On, Tune In, Transcribe: U.N. Develops Radio-Listening Tool. *NPR*. https://www.npr.org/sections/alltechconsidered/2016/10/31/500072478/turn-on-tune-in-transcribe-u-n-develops-radio-listening-tool

Government of Uganda. (2015). *Second National Development Plan (NDPII) 2015/2016 – 2019/2020*. http://npa.go.ug/wp-content/uploads/NDPII-Final.pdf

Government of Uganda. (2017). *The Uganda Population-Based HIV Impact Assessment (UPHIA) 2016–2017*. https://www.afro.who.int/sites/default/files/2017-08/UPHIA%20Uganda%20factsheet.pdf

Hidalgo-Sanchis, P. (2016, October 14). *UNSDG | Tune in to radio to take the pulse on peace and justice in Uganda*. Action 2030 Blog. https://unsdg.un.org/latest/blog/tune-radio-take-pulse-peace-and-justice-uganda,https://unsdg.un.org/latest/blog/tune-radio-take-pulse-peace-and-justice-uganda

Hidalgo-Sanchis, P. (2021). UN Global Pulse: A UN Innovation Initiative with a Multiplier Effect. In M. Lapucci & C. Cattuto (Eds), *Data Science for Social Good: Philanthropy and Social Impact in a Complex World* (pp. 29–40). Springer International Publishing. https://doi.org/10.1007/978-3-030-78985-5_3

Internet World Stats. (2020). *Usage and population statistics*. https://www.internetworldstats.com

Jun, Y. (2022, January 12). *Scaling Radio Analysis with Data Science for Infodemic Monitoring • UN Global Pulse* [UN Global Pulse]. https://www.unglobalpulse.org/2022/01/scaling-radio-analysis-with-data-science-for-infodemic-monitoring/

Kirkpatrick. (2021, May 8). *Speaking of the future: Make sure no voice is left behind | UNESCO Inclusive Policy Lab*. https://en.unesco.org/inclusivepolicylab/analytics/speaking-future-make-sure-no-voice-left-behind

Lemm, S. (2021, December 4). *Data Privacy and Protection Assessments in Radio Mining*. UN Office of Information and Communications Technology. https://unite.un.org/blog/data-privacy-and-protection-assessments-radio-mining

Madamombe, I. (2015). Community radio: A voice for the poor. *Africa Renewal*. https://www.un.org/africarenewal/magazine/july-2005/community-radio-voice-poor

Menon, R., Biswas, A., Armin, S., Quinn, J., & Niesler, T. (2018). Automatic Speech Recognition for Humanitarian Applications in Somali. *Proceedings of SLTU*. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), Gurugram, India.

Menon, R., Saeb, A., Cameron, H., Kibira, W., Quinn, J., & Niesler, T. (2017). Radio-browsing for developmental monitoring in Uganda. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5795–5799. https://doi.org/10.1109/ICASSP.2017.7953267

Ministry of Health of Uganda. (2020). *Uganda Health Marketing Group*. https://www.uhmg.org

National Information Technology Authority (NITA) Uganda. (2018). *National Information technology survey 2017/18 Report*. https://www.nita.go.ug/sites/default/files/publications/National%20IT%20Survey%20April%2010th.pdf

Uganda Communications Commission (UCC). (2018a). *Post, Broadcasting and Telecommunications Market & Industry Q2 Report*. https://www.ucc.co.ug/wp-content/uploads/2017/09/Communication-Sector-Performance-for-the-Quarter-ending-June-2018.pdf

Uganda Communications Commission (UCC). (2018b). *Telecommunications, Broadcasting, and Postal Markets Industry Report Q3—September 2019*. https://uccinfo.blog/2020/01/30/communication-sector-report-september-2019/

UN Global Pulse. (2017). *Using Machine Learning to Analyze Radio Content in Uganda*. https://www.unglobalpulse.org/document/using-machine-learning-to-analyse-radio-content-in-uganda

UN Global Pulse. (2018). *Using Big Data and Machine Learning to Respond to the Refugee Crisis in Uganda*. ReliefWeb. https://reliefweb.int/report/uganda/using-big-data-and-machine-learning-respond-refugee-crisis-uganda

UN Global Pulse. (2019). Ending violence against women and girls in Uganda. *UN Global Pulse*. https://www.unglobalpulse.org/project/ending-violence-against-women-and-girls-in-uganda/

UNESCO. (2014). *Community Radios Promote Local Development across Africa*. https://en.unesco.org/news/community-radios-promote-local-development-across-africa

United Nations Development Programme. (2016). *Uganda Human Development Report 2015: Unlocking the Development Potential of Northern Uganda. Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience*. http://hdr.undp.org/en/content/uganda-human-development-report-2015

United Nations Development Programme. (2018). *Human Development Indices and Indicators: Statistical Update*. http://hdr.undp.org/sites/default/files/2018_human_development_statistical_update.

United Nations Human Rights Council. (2020). *Uganda Comprehensive Refugee Portal*. Available https://ugandarefugees.org/en/country/uga

United Nations Women. (2020). *Global Database on Violence against Women*. https://evaw-global-database.unwomen.org/en/countries/africa/uganda

Uniting to Combat NTDs. (2018). *Reaching a Billion. Ending neglected tropical diseases: A gateway to Universal health coverage*. https://unitingtocombatntds.org/reports/5th-report/

World Bank. (2009). *Human Development in Uganda: Meeting Challenges and Finding Solutions*. World Bank. https://openknowledge.worldbank.org/handle/10986/12658

World Health Organization. (2021). WHO and UN Global Pulse are building a social listening radio tool to aid the COVID-19 infodemic response. *World Health Organization*. https://www.who.int/news/item/04-05-2021-who-and-un-global-pulse-are-building-a-social-listening-radio-tool-to-aid-the-covid-19-infodemic-response

9

Extracting Insights in and around Government

Experiences in South Africa

Vukosi Marivate and Nyalleng Moorosi

9.1 Introduction

With growth in data availability and computing power, governments and public institutions are becoming more aware of the benefits of data-driven decision making (Provost and Fawcett, 2013). Current work in data science has highlighted projects where data science and machine learning have been used to alleviate social problems, influence policy, and overall improve governance. In Africa, we have projects which range from the use of computer vision methods for agriculture and crop disease You et al. (2017) and Pryzant et al. (2017), to traffic collision understanding and prediction of lethality with municipal police (Mokoatle and Marivate (2018)). We have also seen work such as that reported by Steele et al. (2017) and Jean et al. (2016), who used remote sensing technologies and machine learning to understand income disparities in communities and poverty estimation—information that is invaluable in planning and policy development.

In this chapter we introduce important determinants of a successful deployment of data science methods for a government using our experience in South Africa. We discuss key factors which have seen us either succeed or fail. Below we list them and expand in later chapters. We hope that by sharing this work we will get practitioners closer to realizing the benefits of data science for society (Catlett and Ghani, 2015). We close this chapter with what we learned about what happens when the stakes are higher—that is, when this work has to be done during a pandemic.

Vukosi Marivate and Nyalleng Moorosi, *Extracting Insights in and around Government*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press.
© Vukosi Marivate and Nyalleng Moorosi (2025). DOI: 10.1093/oso/9780198879145.003.0010

9.1.1 Determinants of Success in ML for Government Projects

1. **Political will and buy-in.** In engaging with large institutions with multiple stakeholders, an internal champion is essential. This person need not be an executive manager, and in fact in our experience we have found that operational and mid-level managers are the best champions—they have enough decision-making power, but they are also on the ground enough to know what avenues are possible.
2. **Project scoping.** Some government departments have high churn and expectations can change with the department staff. The people in these offices answer to elected officials and very often that elected official will exert pressure on their department to get results which are good for their political careers. As a scientist on the project, it is important to stay within the agreed scope and never over-promise. A well-documented and airtight agreement is essential before any work begins.
3. **Availability of skills and expertise.** Once a research project is complete, it has to leave the research lab and live on government servers and be managed by the department or municipal ICT personnel. When deploying a machine learning model or data science tool, it is important that the ICT department of the institution can handle the maintenance and running of this system. Additionally, given how scarce data science skills were in South Africa when we began, we knew that we had to go beyond training departments we directly worked with.
4. **Understanding current data readiness and accessibility.** Without data, there is not much for data science to solve. That, however, is only the tip of issues related to data readiness. There are matters of privacy and governance, format, ownership, sharing and many others to consider before deploying a successful data science project.

9.2 Data Science for Society in Practice

9.2.1 A Practitioner's Experience of Data Science for Government

As the primary research arm of the South African government, the Council for Scientific and Industrial Research (CSIR) is often tasked with investigating the applicability of new technologies in government processes.

As researchers at the CSIR, we spent a lot of our time exploring the use of machine learning techniques in several government processes and departments.

What is data science to us? Data science is a multidisciplinary field that has arisen in response to increasing availability of data, cheap compute and storage, and that aims to use this data to answer some of the world's challenging questions. The "science" of data science refers to using a rigorous scientific approach to answer these questions. The field tends to be connected with mathematics, statistics and computer science methodologies, extended and invented to work on a wide array of other fields. With the great success of large organizations (especially large internet companies) in leveraging data in order to create new products or solve very challenging problems, there has been a rise in the demand for data science skills in many spheres of society, government included.

In this chapter, we provide an illustration of how the four factors we mentioned in the introduction manifested. We use four past projects for illustration.

9.2.2 The Case of Energy Management in a Large Municipality in South Africa

Due to the high rate of urbanization in South Africa, energy, especially electricity management and allocation, is an ever-present problem for municipalities (Ye 2018). For a large South African municipality, energy modeling in particular is important. That is because beyond the need to understand the national grid usage, electricity sales make up a large portion of the non-tax revenue for municipalities (StatsSA, 2020). In South Africa, analysis of electricity usage data not only brings in an understanding and planning of the electrical needs in the community, but it can also help identify households which may be illegally connected to the grid (electricity theft) and are not paying for electricity consumption (Makhubele, 2020).

As a national research institute, we first aimed to work with municipalities. To start off, the relationship with municipalities was exploratory. One of the first municipalities we worked with had recognized the opportunities to improve their delivery of services to citizens by leveraging the data they collected. Through numerous interactions, ideas were generated on which areas of work would make impact given the internal and external environment of the municipality. We had also identified a number of champions within the municipality who straddled innovation and technology leadership. These were the champions we needed.

To start, we looked at how the demands for power were changing in the municipality over time. This work was neither novel nor unique to South Africa, but it provided a good starting point. First, we looked at electricity buying patterns of households to understand these demand patterns. An interesting find in this project was that combining weather data and purchase data can give us clues on which households used solar for their energy needs (when it was very humid, rainy or with lots of cloud cover, these customers purchased electricity). Furthermore, we could differentiate geographical areas that had constant spending patterns from those with unpredictable patterns. We were also able to detect total shifts in spending patterns—not simple fluctuations.

What was great about working with this team was we found an appropriate champion early in the process. An appropriate champion needs to be passionate about the project and believe in its importance because a lot of their time will be required going forward. It helps if the person has a long history with the organization as they will have a full understanding of external factors (such as political climates) and players who can impact the success of the project.

Finally, and probably most importantly, an appropriate champion is one who is senior enough to mobilize resources but operational enough to fully understand what is feasible. There were projects such as the outcome monitoring project we discuss in the following section, in which we had a very senior champion and what that can do is: 1. Senior staff in government tend to be too close to elected officials who often have to sell a lofty story—therefore not very practical. 2. Sometimes it can be difficult to get buy-in with operational staff as they may resent being dragged into more work by their boss just because some "scientists claim to know how to do part of their work." 3. There is often very high churn with senior staff as they are often hired by the current minister/representative. Also due to this high churn, operational staff may decide to not invest themselves because the instructing individual will probably be gone in the next election cycle or reshuffle.

We met this person after making a pitch at the municipality and with him, we built tools and trained next-generation data scientists.

9.2.3 Project Scoping and Overpromising: The Case of Outcome Monitoring for a Government Department

In 2016, we began a project to monitor outcomes for different interventions funded by the Department of Science and Technology (DST). At the beginning, all parties had lofty goals and ideals they wanted to achieve. We

imagined a completely automated system that could both understand narrative from reports and numerical data from budgeting documents and be able to collate and corroborate the claims in the reports with the numerical data. This was a desired document because it could be one of the ways to highlight inconsistencies between narratives in reports and budget allocations. To begin, the project was given a year to pilot. For the pilot we wanted to extract narrative from budget reports and compare it with the numbers in the budget documents.

In this case we once again had a fairly large institutional buy-in, led by a senior manager in the department. To begin, though there was strong interest from the DST, it was a challenge to get the budget and other resources approved in a reasonable time frame. Thus, as a way to reduce the inertia in starting the projects and to quickly build a proof of concept, we set up this project as a student/intern project. These sorts of student projects were for the most part free for partners because the internship cost was covered by another system, whose main concern was that a student builds a "real" product. This however proved to have been the wrong to go about this type of engagement. The process of building a working proof of concept for free may have eventually led to our pricing problem later on in the project, where the client expected a lot more work from the budget even though the project had moved beyond being an intern project.

Additionally, the initial success of the project (probably also had to do with the fact that it had been free up to now) increased expectations on what could be achieved. When the DST project began, we had a tool that could provide a document summary and provide some intuition on how much each of the reports' narratives aligned with the overall departmental strategy, and temporal characteristics of the narrative. It was from the demonstrations of the document summarization tool that we received requests to extend the document summarization to a budget evaluation and monitoring tool. In a number of iterations, we worked to increase the scope of the tools to cover budget monitoring and evaluation. But as we continued on with the project it became apparent that the problem was not well scoped, not well prioritized and the expectations between us and the DST were not well aligned. Unfortunately, it soon became clear that we could not achieve the scope that the DST wanted with the budget (both time and financial) they were providing. This was a learning moment for both parties. At the end we delivered two systems, one narrative summary and the other a dashboard to visualize the numerical data.

This highlights the delicate balance that has to exist between researchers trying to look at the challenge in front of us as an opportunity to explore

and uncover, while at the same time having to make sure to set up the right precedence for how they wish to continue going forward.

9.2.4 Skills and Expertise: The Case for DSIDE and the Indaba

In any developing technology, human capital is one of the strong determinants. In South Africa, the development of this skills pipeline has been undertaken by universities, government departments and research institutions as well as private industry. In the field of data science one of the most well-known large data projects which required a large number of trained practitioners was the Square Kilometer Array (SKA) project. The SKA, currently being constructed, will be the largest radio telescope in history,[1] and requires advances in many fields such as data processing and data science to achieve its goals (Davidson, 2012). While there were adulations for South Africa for succeeding in hosting this project,[2] it became a sore point that there were not all that many South African scientists who could exploit the collected data (at the time almost all the data was sent to European institutions for analysis) (Mzekandaba, 2015a, b). Thus began the project of building internal capacity.

One of the big components of the SKA human capacity development was a two-week summer school held every winter (June or July) in Cape Town. We built on this model and set up a 12-week immersive program divided into 4 weeks in the winter and 8 weeks in the summer. This we named Data Science for Insight and Decision Enablement (DSIDE). Though we had great success at DSIDE, we also decided to join efforts with other African researchers in leading research institutes and developed a week-long intensive program for just machine learning (this is the mathematical engine that powers data science and AI in general), which we called The Deep Learning Indaba. Below we discuss these.

9.2.4.1 Data Science for Insight and Decision Enablement

DSIDE was a program, funded through the Department of Science and Technology (DST) and implemented at the CSIR. The CSIR recruited 50 to 60 university students to spend three months during their vacation period working on data science problems. The program was initiated in 2014 and focused

[1] Please see: https://www.sarao.ac.za/

[2] Please see: https://www.gov.za/about-government/government-programmes/square-kilometre-array-ska

on projects solicited or submitted by researchers, government departments, NGOs, and private industry. Due to the objectives of the DSIDE program and the nature of the projects, the DSIDE teams were made up of diverse students in terms of level of education, academic discipline, and academic institution. Institution diversity was especially important for us given the stark disparities in education resources in South African universities. In this program we wanted to make sure that students from previously disadvantaged universities received the same exposure as those from well-resourced schools.

In building DSIDE, the emphasis was on team and project project-based learning (Pucher and Lehner, 2011). A typical team had three students, focused on a single challenge for the full term. The program began with a few workshops for students to get acquainted with important data science topics and tools. They were introduced to the command-line in Linux, exploratory data analysis, building data products, project management, and machine learning among other key topics, etc. To start, the students were encouraged to first explore their data, visualize and get a sense for the general trends and types in the dataset before going into building any machine learning methods. This was an approach we wanted to emphasize as at times methods in machine learning can be regarded as more important than understanding one's problem and data (Sambasivan et. el., 2021); an attitude which is actually not correct.

Next, the students were exposed to a number of machine learning methods to choose from—each introduced by a researcher with a solid experience in that method. This allowed students to take advantage of the researchers around them. Additionally, in the 12 weeks of DSIDE, we also hosted external visitors who came and spoke on their experiences on using data. These included government officials, researchers, and private industry professionals.

9.3 Challenges and Wins for DSIDE

The full impact of DSIDE is still to be quantified and will only be apparent years from now. However, in the four years of the program, we have witnessed some challenges and some wins. We begin with the wins.

In South Africa, students at university level tend to be solely focused on their studies unless they are part-time students with full work commitments. Through DSIDE, the CSIR team was able to contribute to the pipeline of building talent for the country by making the students more versed in how to actually implement and solve problems when there is a client with

expectations and deadlines. Students met researchers, considered data science as a long-term career and even enrolled for further study (MSc and PhDs). We worked with students from diverse backgrounds, and for us this not only made for the program to be more interesting for the students (Stahl et al., 2010), but also made it possible for us to learn more about the different programs available across the country. Additionally, the diversity in the teams brought in different types of thinking and formed relationships between students who might otherwise never meet.

To further integrate other ways of working and thinking in data science, we developed a relationship with the Data Science for Social Good (DSSG) program that was started at the University of Chicago. Here we shared resources and co-lectured in one another's programs. Additionally, in 2018 we had our first DSIDE participant who also got into DSSG—a program that is notoriously difficult to get into. That scholar is currently a PhD candidate at MILA in Montreal.

In any outreach program, especially one as transformational as the DSIDE program, one of the first hurdles is recruitment. How do we recruit fairly, how do we allocate the spots to the deserving candidates? In a country as unequal as South Africa, this problem is magnified. We know that one of the determinants of an individual's future success is the wealth in their network (Ray et. el. 2001). Therefore, we knew that as much as we would like to give as many spots to students from previously disadvantaged universities, having a mixture of universities will eventually be more effective. And so, we reached out to our networks to advertise to as many universities as we could reach in the country. From here we used a system that was a mix of student excellence and background. To do this, we had weeks of interviews where we tried to understand the person behind the grades and in several instances we accepted students who were on academic suspension (they failed a subject multiple times and were put on a year or two-year-long suspension). Most of these actually proved to be some of our strongest students and with mentorship from the program they managed to complete their degrees and are now senior data scientists in banks and research institutes.

The second challenge was when we expanded the program from 30 to 60 students, we also needed more technical project leads. To succeed, we wanted to have at least one senior technical lead to oversee a maximum of two projects. Third, similar programs such as the University of Chicago Data Science for Social Good[3] are able to recruit advanced scholars, mostly PhD scholars, but due to a shortage in the number of PhDs in data science related

[3] Please see: https://dssg.uchicago.edu/

fields in South Africa, DSIDE had to begin with undergraduate students. In time we have managed to recruit PhDs, but they are still a super minority.

Reaching 50–60 students a year is good, but the expectations in increasing access to education with experts are much greater. This then required another program to be able to expand the supply of access and mentoring.

9.3.1 Reaching Beyond the CSIR: A Machine Learning Summer School and More

What soon became apparent is that there was a need to build a community and capacity in Machine Learning/Artificial Intelligence/Data Science in South Africa and the African continent as a whole. The Deep Learning Indaba[4] was created in 2017 to fill this gap. As co-organizers, we worked to create a ML/AI/DS summer school that brought together students, academics, researchers, and professionals in an intensive six-day program to strengthen African Machine Learning. In 2017 we had close to 330 participants at the first Indaba, while in 2018 we have expanded to nearly 550 participants. In the first year the focus was to get the program started and to see how we can reach people. In the second year we expanded the content to include more advanced topics as well as allow some flexibility on the last two days for participants to choose topics of discussions in smaller groups—this allowed participants to go into topics deeper than they would in large groups. We added Ethics and Policy in AI, African Machine Learning, and focus sessions on Computer Vision, Natural Language Processing and Reinforcement Learning, just to name a few.

Though the Indaba was a roaring success from year one, we still had a couple of gaps that still need filling:

- How do we support remote faculty? Many universities on the continent have only a small subset of faculty in ML/AI/DS. We wanted to identify ways to support and grow these faculty as they went back to their home institutions. This is still an ongoing problem to this day.
- How do we build mobility of ML/AI/DS professionals and researchers across African countries? How do we get past legacies of colonization where Africa was broken in Anglophone, or Francophone or Portuguese, etc.?

[4] Indaba is a word in the Zulu language spoken in South Africa that stands for a conference or consultation to discuss an important matter.

- How do we facilitate academic supervision for students who do not have access to mentors or advisors familiar with ML/AI/DS in their universities?

We believe the Indaba is a success, not just the annual event, but it has also spawned a community and has also shown that we can come together and use resources to raise each other. Other *Indaba*-like schools are forming across the world in areas not served by the large conferences or with access to as many experts and resources. Examples of these are the Khipu in Latin America and the South East Asia Machine Learning School.

9.3.2 Data Governance and Accessibility: Insights from Working with Municipalities

When collaborating with partners, especially in projects involving data-driven approaches, Data Scientists must ensure that essential queries about data accessibility and management are addressed promptly. Key concerns to explore initially comprise knowing where data resides, who possesses access rights, and how someone can obtain authorization. Addressing these matters helps identify potential challenges that hinder efficient collaboration between parties.

For instance, in our projects with various government agencies, we encountered several instances where acquiring responses to fundamental questions related to data governance proved arduous. Often, there were no established procedures for disseminating information among different departments, resulting in confusion regarding specific roles and duties associated with handling sensitive data. This meant that individuals were a lot more hesitant to share data or any information about the dataset because they did not want to have to be responsible should something go wrong. With robust systems, it is clear to everyone in the pipeline what they can and can't do with data. Moreover, discussions surrounding ethical considerations, accountability, and compliance protocols rarely occurred until later stages of the venture.

To overcome these obstacles, data governance frameworks must exist to outline clear directives governing data management and distribution across an organization. Such measures will allow participants to recognize their roles and responsibilities while reducing misunderstandings and conflicts due to insufficient communication. Applying effective data governance

policies enables more productive and secure collaboration, enhancing overall data quality and availability for informed decisions.

9.4 Conclusion

There is often a distinctive working relationship between government entities and researchers. Research is often a risky multi-year exploration undertaken by individuals who for the most part have job security. Lawmakers and public officials on the other hand are often dictated by expectations and pressures from their constituents, which may require faster delivery times and moving targets. In this chapter we have provided a summary of some of the tricks of the trade we learned through our time at a national lab. We have done this through examples of the work we undertook in collaboration with policymakers with different entities in South Africa. We discussed our efforts in building the data science human resource pipeline with the Department of Science and Technology as well as working with local municipalities in electricity management. This chapter is only a short summary of our work at CSIR, but we hope it has sufficiently highlighted the opportunities as well as challenges when collaborating with the government in data science research.

Acknowledgments

We would like to acknowledge our many collaborators in making the above possible. In particular, we would like to thank the CSIR Modelling and Digital Science Data Science team, students, and management for their support—data science is a team sport. We would also like to thank the DSIDE co-organizers and staff for the wonderful program we have been able to craft together. We lastly want to acknowledge our Deep Learning Indaba co-organizers.

References

Catlett, C., and R. Ghani. Big data for social good. *Big Data*, 3(1): 1, 2015.

Davidson, David B. Potential technological spin-offs from meerkat and the South African square kilometre array bid. *South African Journal of Science*, 108 (1–2): 01–03, 2012.

Forrest, Ray, and Ade Kearns. Social cohesion, social capital and the neighbourhood. *Urban Studies*, 38(12): 2125–2143, 2001.

Provost, Foster, and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1): 51–59, 2013.

Stahl, Günter K., Martha L. Maznevski, Andreas Voigt, and Karsten Jonsen. Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups. *Journal of International Business Studies*, 41 (4): 690–709, 2010.

Steele, Jessica E., Pal Roe Sundsøy, Carla Pezzulo, Victor A. Alegana, Tomas J. Bird, Joshua Blumenstock, Johannes Bjelland, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127): 20160690, 2017.

You, Jiaxuan, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, pp. 4559–4566, 2017.

Mokoatle, Mpho, and Vukosi Marivate. Collision course: Challenges with road traffic accident data in south africa. In 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), pages1–6. IEEE, 2018.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794, 2016.

Ghani, Rayid. Data science for social good and public policy: Examples, opportunities, and challenges. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 3–3. ACM, 2018.

Pryzant, Reid, Stefano Ermon, and David B. Lobell. Monitoring ethiopian wheat fungus with satellite imagery and deep feature learning. In CVPR Workshops, pp. 1524–1532, 2017.

Pucher, Robert, and Martin Lehner. Project based learning in computer science—a review of more than 500 projects. *Procedia-Social and Behavioral Sciences*, 29: 1561–1566, 2011.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. "'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

Simnikiwe Mzekandaba. Lack of data scientists poses SKA problem. ITWeb, August 2015a.

Simnikiwe Mzekandaba. Universities unite to boost SKA. *ITWeb*, September, 2015b.

StatsSA. Quarterly financial statistics of municipalities. Statistics South Africa Reports, 2020.

Makhubele, Wisani. Eskom dismantling illegal connections in diep- sloot. South African Broadcasting Corporation, 2020. URL https://www.sabcnews.com/sabcnews/eskom-dismantling-illegal-connections-in-diepsloot/.

Ye, Yuxiang, Steven F. Koch, and Jiangfeng Zhang. Determinants of household electricity consumption in South Africa. *Energy Economics*, 75, 2018, pp. 120–133, ISSN0140–9883.

PART III

STRATEGIES, TOOLS, AND METHODS FOR INCLUSIVE BIG DATA ANALYSIS IN HUMAN DEVELOPMENT

10
Digital Data Handling

A Tutorial Approach

Eduardo López and Sanna Ojanperä

10.1 Introduction

A prerequisite to carry out data science studies on topics related to development is the creation or acquisition of relevant large datasets and their effective management. Dealing with such data is usually a non-trivial affair. There are ethical, legal, regulatory, infrastructural, and import-export restrictions, as well as a variety of other demands (including personnel) that need to be dealt with and that differ from those associated with smaller and more traditional datasets. While this chapter cannot cover all possible situations, we attempt to give a general enough description of the most common challenges that arise when doing development work with Big Data.

Reflecting the diversity of research on development topics, the data may be location-based or environmental (gathered with sensors placed in a natural area or an urban space), transactional, behavioral, connected to productivity, or based on communication. Data may pertain to the study institutions, the financial system, or people's activities. A common denominator to these data is their large size, enabling the extraction of analytical insight and requiring a new approach in terms of analytical software to manage and process them. Traditional statistical software suites commonly used by development professionals (such as Excel, SAS, SPSS, and Stata) are unfit for the task and need to be replaced with the use of leaner approaches. Adopting basic software development using C (or its derivatives), Python, or R avoids the mixture of weaknesses displayed by software suites under various conditions (some scale poorly with data size, others require as much skill to reprogram as using basic approaches, etc.).[1] Choices about analysis tools can influence how the

[1] The recent explosion of interest on generative AI tools such as ChatGPT, Google's Gemini, and many other Large Language Models may also disrupt this landscape in the near or mid-term future, but its effectiveness is yet to be ascertained.

Eduardo López and Sanna Ojanperä, *Digital Data Handling*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Eduardo López and Sanna Ojanperä (2025). DOI: 10.1093/oso/9780198879145.003.0011

data handling process may be best arranged. Owing to their size, large data sets likely also need specialized storage solutions. Whereas a more traditional dataset used in development work may easily be stored on a computer hard drive, big development data may not be suitably stored on premises and may require the use of co-located data centers or cloud storage.

While this variety of data and associated tools can generate complexities in their management and use, there are in fact a considerable number of similarities across situations. This is because the factors that influence how data are dealt with while maintaining their integrity cut across many data sets. From the technological standpoint, two areas of knowledge deal with the ideas described in this chapter, information security (Whitman & Mattord, 2021) and data engineering (Reis & Housley, 2022). The first addresses how to keep data safe and confidential while the second is more concerned with making data use efficient. Furthermore, other topics pertinent to the material in this chapter include research ethics, protocols, monitoring, and evaluation (OHRP, 2022; TRUST, 2018), legal and regulatory frameworks for data governance and use including export controls and data privacy (related to commerce law, data protection, and data policy), and the use of proprietary data that generate the need for contractual agreements (such as data use and/or non-disclosure agreements that are also part of privacy and commerce laws). Although each and every one of these areas is well covered in research papers, books, reports, and various other sources, a single source collating guidance from them for the purposes of carrying out research on big development data is not readily available.

In this chapter, we attempt to summarize from a purely practical standpoint, highly biased towards the view of the researcher (in either academia or practitioner organizations), how data should be thought of in terms of their acquisition or creation, management (including sharing), and final disposal. We call this sequence of events the **data handling life cycle**. This cycle is affected by some important **data characteristics**: sensitivity of the data, size, associated intellectual property, and legal conditions of use. A final dimension to the data relates to their **use** and **users**: who has access to them and who controls them.

Some large governmental research sponsors such as the US National Science Foundation (NSF) and the research councils under UK Research and Innovation (UKRI), require that researchers present a data management plan when applying for funding and many universities recommend them as part of their research data policies. The kind of information we discuss in this chapter is useful in crafting such plans. Data management plans aid in thinking through how to organize the data, keep them safe, and ensure that the

research can be verified and replicated. They are also useful for flagging up issues related to data management, which may emerge during the research project. While data management plans are a fairly standard part of research on development topics carried out in academic institutions, they are not always required for research conducted by other actors in development such as donor organizations, governmental agencies, or think-tanks and civil society organizations. Developing a data management plan has many benefits in ensuring the integrity of the research involved as well as for ensuring the safety of the subjects of the data, which is why we believe adopting these plans more widely would be of general benefit. The guidance outlined in this chapter may be helpful for this purpose.

Although the most secure way to safeguard data is to treat them as highly sensitive (such as containing personally identifiable information), such a level of stringency is not always necessary and may even be counterproductive. Legal obligations, ethical considerations, and common sense need to be applied. Researchers must strike a balance between adding redundant safeguards and maintaining efficient data usability. In practice, the range of possibilities are large. In some lucky situations, data have a minimum level of sensitivity and are of relatively small size, and therefore can be managed in very simple ways such as storing them on a laptop, with nothing more than password protection. However, as the size and sensitivity of the data increase, a more thought-out approach becomes advantageous or even necessary.

One more dimension to take into account concerns the situation where data creation, a possible part of the data handling life cycle, is involved. In the context of development where it is common to study the problems of underserved, disadvantaged, and vulnerable populations, consideration should be given to the very factors that contribute to these inequalities in order to make data more representative. While this chapter is not aimed at discussing these factors, we take the opportunity to emphasize that studies carried out on topics related to development should also ensure that the project provides benefits of participation to involved stakeholders, avoids diverting local resources, and responds to the needs of the context where research is carried out.[2] Part IV of this edited volume discusses the broader topic of datafication and its intersection with social justice and data governance for those interested in delving deeper into these topics.

This chapter is written without any pretenses to be a treatise in the management of data for research in human development. It is more akin to a practical

[2] See for instance the Global Code of Conduct for Research in Resource-Poor Settings for key principles to consider (TRUST, 2018).

collection of notions to provide researchers with hands-on advice on how to approach data projects and, perhaps, to spare them from some unnecessary mistakes and delays. Keep in mind that, in most cases, the kinds of practicalities discussed here have to be dealt with by researchers themselves, and therefore, it is best to be ready for them.

10.2 Outline of the Data Handling Process

The major aspects of digital data handling mentioned above (life cycle, data characteristics, and data users and uses) interact by affecting how each of the processes of data handling should take place. Concretely, they lead to a **data handling process**. This process does not need to be written or spelled out formally to exist. In very uncomplicated situations, for instance, this management process happens on its own when a researcher simply places some files in a folder of their computer, analyzes them, saves preliminary results, etc., until the research is complete. But in cases when the data are large, sensitive, have major legal conditions of use, etc., such data handling process may indeed require careful analysis and even the need to be stipulated "on paper".

For organizations that expect to work with data frequently, it may be advisable to develop institutional guidance and governance processes tailored to the specific remit of the work carried out, along with regular reviews of those processes. Central areas to consider include data architecture and storage, data governance, data quality assurance, and data quality control.

To give a more complete sense for how the process emerges, we consider in some detail the data aspects below. After doing so, we present a complete test case in which these aspects come together in a particular data handling process.

10.2.1 The Data Handling Life Cycle

There is a sequence of events that data used for research or study purposes generally follow. This are diagrammed in Figure 10.1. While there is variety in the way data are acquired, once they are in the possession of the researcher, they typically follow the same path. Before we consider each of these general phases, it is useful to describe them briefly.

The first stage of data handling is the actual data acquisition. This can take a variety of forms ranging from primary data being generated by the

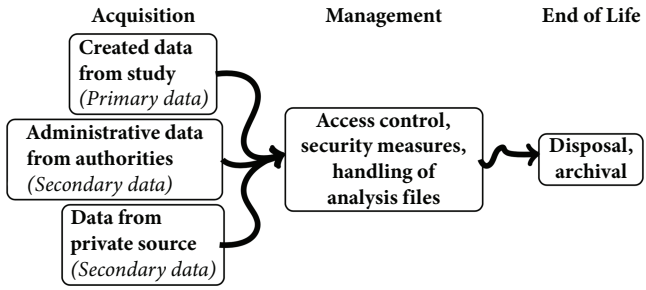


Figure 10.1 General schematic of the data handling life cycle.

researchers for the specific purpose of doing a particular study to the use of secondary data, such as large administrative data generated by third parties (e.g., governments), for a broad set of possible lines of research.

The second stage of data handling is data management or maintenance. This task is not simply limited to maintaining the data under the same conditions and location where they were first placed when acquired. In fact, data analysis is a highly dynamic process that generates new data and intermediate outputs before the final results of the study are achieved. It is therefore important to establish some clear ground rules about how data are to be handled along the duration of a project, and whether they will be stored or archived for a certain period after the study to enable reproductive research.

The final stage involves the end of life of the data, consisting of the disposal/destruction of the data as well as archiving of allowed data for possible future verification of old results or other purposes. This step requires some care. In general, data obtained under certain conditions, such as a use license, may have explicit rules about how to dispose of them. In general, the most destructive method available to you is the best one. Even if a data-use agreement does not require disposal, a research team that has stopped using a data set with sensitive information should carefully consider the risks of keeping such data. Once the decision for data disposal is taken, methods such as physical destruction tend to be the best. The ultimate goal is to make the data irretrievable. Any surviving archival data, if possible, based on the situation, should be archived according to agreements and/or ethical considerations: any sensitive elements of the data need to be treated with the same care as in the original data, and in effect elicit a kind of reassessment of the data as if they were new data. Archival data that can be released should be modified so that no sensitive data can be extracted.

10.2.2 Data Characteristics

A different dimension to the data handling life cycle is given by the characteristics of the data. For us, characteristics refer to features attached to the data itself. Concrete examples of data characteristics include sensitivity of the data, size, intellectual property, and possible legal conditions for use and analysis of the data. Data characteristics have a major effect on the way each of the elements of the life cycle are tackled.

All data characteristics play a major role in how data are handled, together with the details of use and users discussed below. For instance, the size of the data set has a direct bearing on the kind of infrastructure you need to use to handle the data. Small data sets can, as mentioned above, potentially be dealt with by using devices like a laptop computer (although sensitive data may be less well-served by this approach); large data sets may require the use of large, dedicated servers. Backup processes may also relate to data size, as small datasets can be backed up easily, but large ones require more complex infrastructures and rules. There may be a need to use co-located data centers or cloud storage solutions (but consider the legal ramifications discussed below).

Expanding on the consequences of size, it is generally possible with a little bit of information, to estimate how much storage is needed for data based on their details. Let us consider a data set that tracks n units (say individuals or households), each one described by m features (for example income, education, and other characteristics), over t points in time. Assume further that each such record requires q bytes of data and that, as a general rule, no less than a factor of two needs to be applied because during data analysis (or backup), the amount of data duplicates in size; in fact, this factor tends to be considerably larger (in some cases 10 or more), although when data are extremely large and any analytical use is highly managed, the factor of two tends to be realistic. Then, the storage needed for the data is of the order of

$$\text{Storage} \;\geq\; 2 \times n \times m \times t \times q \;\text{ bytes} \tag{1}$$

Sensitivity associated with the data due to the subject of the data (e.g., personal information, financial information) affects the sorts of arrangements one needs to put in place. For example, publicly available statistical data that a researcher can obtain require little or no security. At the other extreme, if a data set has specific information about individuals, institutions, very well-resolved locations, etc., strict privacy and security procedures such as

strong firewalls, limited data access, encryption, and other methods may be necessary. These, in turn, may need to rely on types of software and/or system/network management not readily available in general.

In many cases, there are legal and ethical processes that come with data characteristics. For instance, when data have been obtained at great expense and may be the product of a process under intellectual property, or are owned by a private entity, data use agreements or data licenses may need to be put in place. These agreements are legal documents that spell out what can be done with the data, how to store them, and how to destroy them. These agreements are typically entered into between the owners or managers of the data and the institution that employs the researchers. Note that, in relation to the comments above regarding the use of data centers or cloud storage, one must be aware that legal or contractual requirements may limit which of these solutions can be taken. For example, some data cannot be exported due to legal constraints, which means that use of cloud storage could be in violation of those rules (many data banks that support cloud storage, offered by large companies, can be located anywhere in the world and therefore placing data with export restrictions in those banks can violate laws or contractual agreements).

Another situation arises when sensitive personal data is created by researchers for the purposes of a study. In these situations, it is common to put in place consent agreements with study participants which also indicate what the data can be used for. In most circumstances, the information that is planned to be collected in the study is reviewed beforehand by a board that decides about the ethical nature of the data collection. Many universities and research institutes adhere to detailed policies concerning the ethical management of research data and, for instance, universities in the United States have local Institutional Review Boards (IRBs) while British universities have Research Ethics Committees (RECs) often on the departmental and divisional level which are tasked with reviewing and protecting human subjects of research. Internationally, ethics committees at universities and in ministries of health (or other pertinent government agencies) may be tasked with similar responsibilities. The work of these organizations is shaped by the wider legal environment related to data protection such as the General Data Protection Regulation (GDPR) in Europe. Data collection on human subjects requires these reviews to take place before the collection begins. When collecting data in a context of development, one should keep in mind that the policies and laws that govern data protection and their implementation in the country may be more lax. Similarly, where

data collection takes place outside of academic institutions such as within donor organizations, the processes of ethical review may not exist or be as stringent as within academia. Researchers in charge of handling data in such contexts should adhere to the highest standards of protecting sensitive data in order to adequately safeguard the subjects of that data. Useful information is provided by the Office for Human Research Protections (2022). More broadly, possible legal requirements necessary to handle data include export control, Data Use Agreements, and Ethical Research Review (e.g., IRB) (see Table 10.1).

10.2.3 Data Uses and Users

In a project, data may need to be accessible to as few as one person or as many as an entire research team, including systems administrators. These two extremes require different approaches to data handling. When a single individual uses data, the rules are simple: a single individual will have access to the entirety of the data, as well as any derivative files created in the process of analyzing them. On the other hand, if a team of people are all meant to have access to a data set, there is likely a need to manage the extent of that access. Issues that may arise relate (but are not limited) to granting or withdrawing access, allowing subgroups of people to have visibility over each other's work or suppressing that visibility, managing partial allocation of data to individuals, etc.

The nature of the use of the data itself is another important aspect. Will the data be used to generate additional information such as intermediate data sets, cuts of data for public release, visualizations, etc.? All such questions need to be considered. In circumstances when the data them sensitive, it is common to have restrictions on what can be published, a restriction that is enforced in practice by review procedures conducted by a review team. The review can be aimed at ensuring that no sensitive information is released, that information cannot be reconstructed from the publication of analysis, or that intellectual property associated with the data is not lost or jeopardized.

Table 10.1 A list of possible legal requirements necessary to handle data.

Possible legal requirements to handle data	• Export control • Data Use Agreement • Ethical Research Review (e.g., IRB)

10.2.4 The Handling Process

Data characteristics and users feed into the way that each stage of the data handling life cycle needs to be approached. First, at the stage of data acquisition, all aspects of the data characteristics must be addressed: can the researcher's infrastructure handle the size of the data? Does the researcher have the capacity to gather, purchase, etc., the data? Similarly, identifying the users of the data and how they can access them is important even at this early stage. Are the users in the same location? Do they have access to different infrastructures? Large digital data can be stored and accessed in various formats ranging from on-premises servers to data centers and cloud services, the choice of which impacts the ease of access and level of security and protection from unauthorized users. These are particularly pertinent considerations given that remote working arrangements have become more popular since the COVID-19 pandemic (Babapour Chafi et al., 2022). Once a thorough assessment has been completed, those responsible for procuring the data can simultaneously put in place the infrastructure, contracts, and rules that will support the data use throughout the life of the project.

If the first stage of the process has been well-managed, the second stage of the data life cycle can be tackled more smoothly. Ideally, the infrastructure in place is likely to be sufficient to maintain data analysis capabilities for the duration of the project. The rules applied to the initial users of the data can be enforced for any new users. But a new component to the process enters the picture: how to handle the generally copious amount of information generated from data analysis? This will depend on the level of sensitivity of such analyses and the original size of the data. It is possible that sensitivity could increase in analyzed files, which may necessitate a further revision of the rules of use. Sharing files between project collaborators is another aspect that may require some guidelines on the basis of the type of analysis and details associated with the project. Yet, another aspect to consider is the tracking and/or containment of generated data. This is important for the last stage of the project, when data must be disposed of and knowing all possible copies of them becomes critical. In addition, if a research project involves publication, the parameters of what is acceptable for being made public must be defined and be consistent with any obligations stemming from any data-acquisition contracts or sensitivity of the data. Many journals and other publishers require enabling access to the data underlying the research contribution for a certain period of time in order to allow for reproducible research, which will additionally impact the arrangements in place for data storage and shareability. Where the data concern or are collected with the

help of research participants, it is salient to ensure that participants and their communities have access to research outcomes.

The final, end of life stage may partly be dictated by the conditions under which the data were acquired. In many situations, contractual obligations dictate how data need to be disposed of. In that case, it is imperative that these rules are followed in order to prevent any contractual breaches. If multiple copies of data exist, including the multiplicity of intermediate analytical files, it is likely that all such information will need to be handled in the same way as the original data. Archiving of data, when allowed, requires assessment again, especially if the data can be archived with sensitive elements; if sensitive elements are not allowed, then archival data should be modified to achieve this goal.

10.2.5 A Summary: Heuristic Approach to Data Handling

The considerations laid out above lead to a basic heuristic framework we now explain. This may not fully apply to everyone's data and therefore readers should maintain a critical posture when addressing their own situation. We bypass the critical question of *how* a team manages to establish the relationship that leads to data, as this is not about data handling procedures.

Perhaps the hardest and most time-consuming step is in the initial stage when both the data acquisition and data assessment take place. Much of what happens at later stages of data handling stem from the way in which the first stages of handling have occurred. Let us dive into the details.

Step 1: Initial Data Acquisition and Data Characteristics Assessment. We combine these two together because data acquisition in complicated situations cannot actually happen without a level of data assessment. A team must first understand what is in the data to identify sensitivity, legalities, size, users, and uses. In doing so, several tasks will emerge:

1. Determining the details of any contractual agreements (when acquiring existing data) or research protocols (when generating new data) and putting these in place.
2. Procuring data storage to the appropriate capacity and level of security.
3. Ensuring that users and uses are considered in the above two steps, including the need for any training or certification of the personnel involved in the project.
4. Transferring and taking possession of the data, which can be a nontrivial task when data are highly sensitive, very large, or both—in some

cases special provisions may need to be put in place, like the shipping of a physical hard drive.
5. Establishing a data management plan that stipulates how data will be handled going forward.

Step 2: Data Management. At this stage, when the data are now accessible to key personnel of a team, other team personnel will begin to work on the project. It has already been indicated above that personnel may require certification (e.g., on how to handle personally sensitive information), and there may be conditions on the person's employment, nationality, etc. Large projects often have some personnel rotation over time, which also means new personnel need to be certified and verified properly and departing personnel should provide assurances that they no longer have access to data. Some concrete tasks that emerge at this stage include:

1. Providing personnel with responsibilities over the data a clear set of rules by which to operate, for example by ensuring they have the data management plan.
2. Ensuring incoming personnel have the proper certifications or satisfy legal conditions for data access.
3. Ensuring outgoing personnel provide sufficient assurances that they will no longer have access to data once they leave the project.
4. Providing actual data access to personnel, including appropriate conditions for them to work on the data (including the data storage and software), that also maintains data security.
5. When data products such as articles, policy recommendations, or other results stemming from analysis of the data are to be made public, ensuring that the material release complies with any contractual obligations established in Step 1 above.
6. Securing proper and secure backup of the data.

Step 3: Data End of Life (Archiving and Disposal). In most cases, sensitive data require procedural disposal, potentially within a given time frame. Part of the data that are deemed non-sensitive and necessary for research assessment, auditing and reproducibility are to be kept in compliance of contractual obligations as mentioned in Step 1. In concrete terms, this leads to the following tasks:

1. Original data as well as data generated in the analysis process should be disposed of under the conditions established in Step 1 or, absent

Table 10.2 Matrix to record the needs of a data research project.

		Acquisition	Management	End of Life
Data Characteristics	Sensitivity Size Legalities			
Data Use	Users Uses			

such agreements, in the most destructive way that guarantees sensitive information will not be released.

2. Data that can be legally archived are handled according to the conditions set in Step 1 and/or to ensure they will safeguard sensitive information.

The discussion above suggests a matrix through which one can plan how data will be handled in a particular project. This matrix can be found in Table 10.2. To make use of this matrix, one should explore how the data characteristics and use affect each of the life cycle elements. The information that is most useful to input are the handling requirements pertinent to each of the rows and columns. For example, if the size of the data affects their management (by determining the minimum amount of storage needed after application of Eq. 1), this can be inputted into the matrix. By the end of the process, information about how to handle the data will be easy to find and act on.

10.3 A Case Study in Data Handling

Consider the following hypothetical situation: a study is set up to evaluate the dynamism of local economies in low- and middle-income countries.[3] The data to be used in this study come from a large mobile phone company that offers mobile money, a service by which people utilize the phone company effectively as a banking entity and their phone as a form of debit card. The data come large, covering millions of subscribers to the service over a period of a year. The number of transactions per individual may be of the order of 20 to 40 transactions per month. The team that will engage in this study includes researchers based at a university in a different country than the subscribers for the mobile money service, and is comprised of a lead

[3] We use the term low- and middle-income countries to refer to the countries that fall into the economic classification of low-income, lower-middle-income, and upper-middle-income as defined by the World Bank (2023). Please see the Introduction for a more detailed discussion about why we chose to use this term.

investigator, employed researchers, and graduate students hired as research assistants. The team also includes researchers local to the country of the subscribers who work in an NGO focusing on financial inclusion. The goal of the researchers is two-fold: write academic papers about the nature of the local economies studied and generate practical advice for policies to put in place to make the flow of transactions more dynamic, leading to better development outcomes for the users of mobile money services. What sorts of arrangements may need to be put in place to perform this data-driven and multi-stakeholder project?

The situation described above has the typical elements of projects with large, sensitive data. Let us now apply the above discussion to the situation.

1) First, what are the data characteristics?
   (a) Data sensitivity: the data are highly sensitive, combining personal and financial information.
   (b) Data size: without need to specify many details, it is fairly easy to estimate the level of data storage necessary. With a country where there are millions of users of the service, we may be in the range of 1 to 10 million users. In addition, 20 to 40 transactions a month translate into between 240 to 480 a year. With one year of data, and only focusing on the lower bound of a million subscribers, there would be 240 to 480 million records. If a single record contains information such as monetary amount of transaction, date, and recipient, this could reasonably require a kilobyte of ASCII-encoded data. This would lead to between 240 to 480 gigabytes of space in an extremely compact way. Following Eq. 1, any analysis files can generate data of comparable size (a minimum factor of two is a prudent, conservative estimate) and therefore it is possible that a minimum capacity of data handling could require at a minimum one terabyte of space, or more generally in the order of several terabytes. If backups of the data are planned, which are always highly recommended, the additional space needs to be considered as well.
   (c) Legal terms of use: data obtained from a private source such as a mobile phone company will very likely require contractual non-disclosure agreements. Such agreements could contain, among others, details about how to handle the data and share them among members of the research team, how to clear material to be published, and how to destroy the data.

A summary of the information can be found in Table 10.3.

Table 10.3 Key acquisition aspects of study of mobile money.

		Acquisition
Data Characteristics	High sensitivity: personally identifiable data including financial information	Secure environment with restricted access
	Large size: of the order of 1–10 TB	Large computational storage capacity and processing
	Legalities: Non-disclosure agreement	Negotiate language of agreement, clarify rules. In this case, consider export control rules when drafting the agreement
Data Use	Users: Researchers and IT administrators in different countries	Contractual, written agreements by team members to abide by rules. Also, any certifications needed to handle such sensitive data. In this case, due to international collaboration, ensure export control rules are clear
	Uses: Academic publications and policy advice	As part of a non-disclosure agreement, spell out rules to make data public, who can see the data, and how to handle data at end of cycle (disposal and archive)

2) Who will use the data and what are the aims of the research project?
   (a) Users: as indicated above, a team of researchers would participate. Note that the participation of collaborators from a local NGO has an impact on legal arrangements. This is because commonly it is the institution of employment of the lead researcher that enters into a data-use agreement. In cases such as this, one may find several modalities of arrangements. Any external collaborators (in our example, members of the NGO) may be granted temporary positions at the institution that has signed a non-disclosure agreement. In such an arrangement, the burden is in creating such temporary positions (e.g., affiliate or visiting researchers) for the external collaborators in the lead organization that has signed the data-use agreement. Alternatively, the institution of any external collaborator may enter into an agreement with the institution initially receiving the data (provided this is allowed under their data-use agreement). Another possibility is that the institution of

the external collaborators signs a direct data-use agreement with the data owners. In any case, this legal requirement is likely to take some time to resolve. Beyond the research team access, it is very likely that systems administrators will be involved to handle the computational infrastructure needed for this project; their legal access to the data must also be arranged.

(b) Uses: since the data will be used for both academic publications and policy recommendations, information about the population represented in the data will become public in some way. It is customary to find two arrangements related to this. First, data reported takes a form by which is it not possible to "re-identify" an individual. This is often achieved by aggregating information or reporting on characteristics of individuals but not the individuals themselves. A second arrangement that helps protect privacy is the review of materials by another party prior to the release of information. Quite often, in this case, the data owner would be the mobile phone company that has shared the data with the research team.

A summary of the information related to data management can be found in Table 10.4.

Table 10.4 Key data management aspects of study of mobile money.

		Management
Data Characteristics	High sensitivity: personally identifiable including financial information	Periodic review of secure environment with restricted access, inventory data.
	Large size: of the order of 1–10 TB	Monitor evolving needs based on data analysis. Monitor location of new analysis files
	Legalities: Non-disclosure agreement	In case agreement generates challenges, discuss with data originator
Data Use	Users: Researchers and IT administrators in different countries	Continue to manage data access. Handle new team members, exit of old members. For new members, check they are cognizant and compliant with use rules.
	Uses: Academic publications and policy advice	Monitor that rules about use and release of data are being followed

The details associated with these data dictate the management approach needed. Let us highlight the critical aspects:

a. **Legal arrangements:** Data owners (mobile phone company), leading members of the research team, and the legal and research services of the various institutions will discuss the data sharing agreement. Such discussions can take from weeks to months.
b. **Infrastructural arrangements:** Because sensitivity requires safeguards from unauthorized access and use, data would require space for storage and processing that has authentication protocols in place. In the most sensitive cases, additional safeguards could be needed such as encryption, or even a system that has no access to the outside world.
c. **Protocols to handle data analysis, creation, and destruction:** These protocols would ensure that, for example, teams of people working on a limited part of the project do not have access to data on other parts of the project unless there are changes in the scope of the work done by individual team members. For example, if the mobile data are being looked at to determine patterns of communication and separately patterns of spending, it may be possible to keep the work of these teams apart and separate the data accordingly. New data created from the analysis of the originally obtained data need to be stored according to agreed-upon protocols. At the end of the project, when the data are no longer needed, or when the contractual agreements with the data owner end, a plan to destroy the data has to be put in place. For example, if the data have been stored in hard drives, these could be physically destroyed in order to prevent undesired data disclosure.

A summary of the information related to the end of life arrangements (of the study outlined in Table 10.4) can be found in Table 10.5.

10.4 Summary

Arrangements to handle large data must be looked at in the light of a collection of aspects that require careful tracking. While various theoretical frameworks can be drawn upon to understand our heuristic (data policies, data engineering, contract and/or privacy laws, etc.), it is ultimately a practical matter. Those engaged in this sort of work will go through the tasks of handling large sensitive data better as a function of how efficiently they can navigate the necessary arrangements.

Table 10.5 Aspects related to managing the end of life of data.

		End of Life
Data Characteristics	High Sensitivity: personally identifiable including financial information	For data archiving, when appropriate and allowed (or mandated), ensure no sensitive information can accidentally be released
	Large size: of the order of 1–10 TB	Reassess needed infrastructure for data archival
	Legalities: Non-disclosure agreement	Abide by disposal rules on agreement or dispose of sensitive data by most destructive means
Data Use	Users: Researchers and IT administrators in different countries	Check users properly delete analysis copies of data and maintain any archival information necessary.
	Uses: Academic publications and policy advice	If there are archival data, they should be openly accessible so that others can reproduce results.

In this chapter, we have discussed the various aspects that arise in the process of handling data from beginning to end. The aspects we have covered are generally not collected in a single article or document and in many cases are discovered by researchers over time, as they deal with the many practicalities that arise in preparing and executing research with data (in our case, typically of considerably large size). Effectively navigating the challenges accompanying such projects will provide any research an advantage in terms of time savings, compliance, data use efficiency, and data safety. In the future, we believe it would be advantageous to create training and certifications for researchers in the combination of topics discussed here.

References

Babapour Chafi, M., Hultberg, A., & Bozic Yams, N. (2022). Post-pandemic office work: Perceived challenges and opportunities for a sustainable work environment. *Sustainability*, *14*(1), Article 294. https://doi.org/10.3390/su14010294

OHRP. (2022). *International Compilation of Human Research Standards*. Office for Human Research Protections (OHRP). https://www.hhs.gov/ohrp/international/compilation-human-research-standards/index.html

Reis, J., & Housley, M. (2022). *Fundamentals of Data Engineering: Plan and Build Robust Data Systems*. O'Reilly Media, Incorporated. https://books.google.com/books?id=Z_TFzgEACAAJ

TRUST. (2018). *Global Code of Conduct for Research in Resource-Poor Settings*. https://www.globalcodeofconduct.org/

Whitman, M. E., & Mattord, H. J. (2021). *Principles of Information Security*. Cengage Learning.

11

Sources of Big Data in Global Development

Eduardo López and Sanna Ojanperä

11.1 Introduction

To understand the use of big data for development, it is important to have a concrete and practical sense for what such data are. Therefore, in this chapter, we take a quick tour through various types of data sources thought of as big data that have contributed to our understanding of problems in human development. In doing this, we explain some basic features of the data and some of the ways in which they have been used. We also present a short discussion of big data as a mindset, i.e., a way to conceive of data, which is a fundamental ingredient of the Data Revolution. In addition, as part of framing such data, we briefly discuss shortcomings of big data. This is done not in an attempt to review such problems, but rather as a reminder that skepticism also needs to be part of one's use of big data.

11.2 Big Data as Both Data and Context: The Data Revolution

11.2.1 Data and Mindset

As described in the introduction of this edited volume, as well as in abundant prior scholarly work (see, e.g., Kitchin, 2014), the concept of big data has been defined in several ways. Many of these definitions acknowledge their size, the fact that in some cases they are exhaustive, and that they are being created in real-time, among other definitions. Authors have also identified a mindset connected to the big data movement, fueled by the characteristics of those data, the new results and interpretations that stem from their study, and the methodologies necessary to tackle their analysis.

Eduardo López and Sanna Ojanperä, *Sources of Big Data in Global Development*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Eduardo López and Sanna Ojanperä (2025). DOI: 10.1093/oso/9780198879145.003.0012

While a high-level, theoretical discussion would be interesting, we take a practical approach here by briefly describing some of the data examples that illustrate big data in several of its most notorious forms. Many of these types of data have, in practice, become synonymous with the notion of big data. Examples of types of big data include earth observation data, mobile phone records, data from the Internet and digital platforms such as social media, and complete administrative records, to name a few. In this chapter, we describe these and other key examples, and identify how their analysis requires new or updated tools in statistics. Many of the data sources we describe in fact overlap with each other, but for simplicity we aim to describe the central characteristics of each data source, their role in development work and research, and point out example applications.

11.2.2 Challenges with Sampling, Representativeness, Quality, Causality, and Algorithmic Bias in Big Data

An important caveat related to using data from some of the sources we will discuss is the appropriateness of the sample. While some of the data are comprehensive, others are sampled with biases, disproportionately capturing certain segments of the population. An important line of work in demography is emerging to address these issues (Cesare et al., 2018). Furthermore, issues of temporal correlation, high dimensionality (i.e., many variables), and several other complications mean that care must be taken when analyzing and interpreting big data (Fan et al., 2014).

As mentioned in other parts of this volume, in many cases big data have not been collected for the purposes for which they are subsequently used. As has been pointed out (Kitchin, 2014), this can have problematic consequences. Although one would like to believe that most sources of big data are indeed comprehensive, this is in fact not the case. A more accurate characterization of data collection is to say it is idiosyncratic, with combinations of information that usually only partially address interesting questions. One could cite many other reasons why big data need to be looked at with a level of care.

The possible shortcomings of big data may be captured through a notion of quality. Intuitively, quality assesses how well some large data source is able to provide accurate and useful information about a situation of interest. A variety of factors come into play when considering this quality, including the representativeness mentioned above, completeness (extent to which the expected attributes are recorded), accuracy (how closely the data measure

relevant attributes), availability (the extent to which data are accessible), consistency (the extent to which the method used to collect the data remains unchanged), and timeliness (how relevant and recent in relation to the phenomenon of interest the recorded data are). From the standpoint of these dimensions, some sources of big data may be better than others, but unequivocal assessments are hard to perform.

In the research related to social science perspectives, the question of whether causality can be inferred from big data plays an important role. This comes across when questions of confounding factors emerge that may obscure the reasons why something occurs (see e.g., Einav & Levin, 2014). However, factors such as this do not mean that big data are not useful, since the patterns that emerge from their analysis can be quite stable, supporting the presence of pertinent information to be extracted.

A more critical and pernicious issue with sample problems is algorithmic bias. Although this chapter is not directly concerned with methods (see Chapter 12 for additional discussion), algorithmic bias starts with the data. Bias in data sampling, when not accounted for properly, can lead to analysis that lacks correct representativeness. The literature on negative effects of data on fairness and equity is growing, pointing out many of these problems, but the momentum driving data use is even greater, risking the entrenchment of these problems (Dalton & Thatcher, 2014; O'Neil, 2016; Noble, 2018; Benjamin, 2019; D'ignazio & Klein, 2020; Graham & Dittus, 2022). Thus, in considering the use of data, understanding of sampling limitations is critical if the data are indeed going to serve the greater goal of development.

The bottom line is this: despite the difficulties and drawbacks, one must be clear there are indeed good reasons to engage with big data, but also to remain aware of their limitations.

11.3 Sources

We now discuss some concrete versions of big data, along with a few examples of their use cases. Some of these use cases have been captured in Table 11.1, which is based on a report by the United Nations Development Group (2017), while others have been found in the literature.[1] However, the examples presented here should not be understood as a representative sample of

[1] We note that the same table is presented in Chapter 12; there a column referring to the methods used to analyze the data can be found, whereas here that column is replaced by one that makes reference to the information source.

Table 11.1 Examples of Research Relating to the Sustainable Development Goals.

SDG	Theme	Example Research	Data	References
1	No poverty	Spending patterns as proxies for income	Credit card or mobile money	(Decuyper et al., 2014; Blumenstock et al., 2015)
2	No hunger	Food-price tracking	Crowdsourced or online tracking	(Kim et al., 2017; Zeug et al., 2017)
3	Health	People's movements to predict the spread of infectious disease	Mobile phones	(Brownstein et al., 2008; Eagle et al., 2009; Salathé et al., 2012)
4	Education	Drivers of student drop-out	Crowdsourced	(Marquez-Vera et al., 2016; Pérez et al., 2018)
5	Gender equity	Spending and financial activity by gender	Credit card and mobile money	(Martinez et al., 2016; Aron, 2018)
6	Sanitation	Access to clean water	Sensors	(Koehler et al., 2015)
7	Clean Energy	Optimization of production and distribution	Smart meters	(Shupler et al., 2021)
8	Decent Work	Micro-level activity like spending, remittances	Postal traffic	(Ansón & Helble, 2013; Hristova et al., 2016)
9	Infrastructure	Public Transport quality and improvement	GPS data	(UN Global Pulse, 2017)
10	Reduce Inequality	Tracking patterns of discrimination	Speech-to-text analytics from mass media	(Menon et al., 2017, 2018, Chapter 8 by Hidalgo-Sanchis)
11	Sustainable Communities	Land encroachment on public land	Satellite images and data	(Bratley & Ghoneim, 2018; Patel & Soni, 2022)
12	Responsible Consumption	Transition to efficient products	Online search patterns and e-commerce	(Zhang & He, 2021)
13	Climate action	Tracking deforestation	Satellite and crowdsource data	(Achard et al., 2002; Hansen et al., 2013)
14	Marine life	Tracking illegal fishing	Satellite data	(Kurekin et al., 2019)
15	Life on land	Tracking on-land events in real time	Social media	(Kibanov et al., 2017)

SDG	Theme	Example Research	Data	References
16	Justice and Institutions	Tracking governance, rights, public opinion	Social media	(Chapter 8 by Hidalgo-Sanchis)
17	Partnerships for Goals	Connecting data for real-time holistic picture	Satellite data	(K. Anderson et al., 2017)

Note: The table is partly an expansion of the table in the guidance note by the United Nations Development Group (2017).

these studies. In fact, new uses for big data in development are being found very often. Here, we only aim to provide an illustration.

11.3.1 Mobile Phone Data: Communication, Mobile Money, and Mobility Data from Smartphones

Mobile phone metadata has become one of the most revolutionary sources of new big data for social research, including the study of indicators like poverty (Blumenstock et al., 2015), highly relevant for the purposes of development. The use of such data cuts across issues of social structure, location, and even financial status, making it both incredibly rich and sensitive.

Structurally, these data are comprised of transaction records from communication through mobile phones, sometimes accompanied by location data, which capture when a phone user contacts somebody else and, in the case of a phone call, for how long they communicate. Combined with the account holder's metadata, mobile phone records can provide a wealth of information about social interaction by the phone user, traces of which social group the person belongs to, personal levels of activity, to name a few. Combining metadata and phone records, socio-economic information can also be inferred. In Chapter 2 of this volume, Oliver discusses projects that approach natural disaster preparedness and crime detection through the use of big mobile data.

When the phones are "smart," additional data may become available such as high temporal frequency location, making it possible to pinpoint people's position and movement patterns through space and time with a high degree of accuracy. In some cases, smartphone data can be paired with other data generated by usage of apps to produce behavioral profiles of people.

The data are usually held by mobile operators (as opposed to, say, governmental authorities), and typical analysis is undertaken for large samples of

phone records, normally at national levels (driven by the service areas covered by the mobile carrier), for some chosen time window. The relatively high frequency of data points generated by each individual, combined with the fact that the number of subscribers to the mobile carrier tends to be in the millions, makes these data sets also very large. Although a large portion of the scientific and applied analysis that has been carried out on such data draws from various established statistical techniques, other methods such as network science and machine learning are being applied.

Scientific studies in high-income countries have mainly focused on social and geographic questions. The use of such data for the purposes of development, beyond questions of poverty, has touched on issues as diverse as population and demographic studies (Deville et al., 2014), energy use (Bogomolov et al., 2016), employment (Toole et al., 2015), improvement of official statistics (Jahani et al., 2017), loan repayment (Björkegren & Grissen, 2019), and food security (Liang et al., 2020), among others. Some research has been driven by ambitious competitions and challenges such as Orange's D4D (Data for Development) Challenge (de Montjoye et al., 2014). Within the non-profit sector, Flowminder (2023) focuses on big mobile data to improve health and welfare in low- and middle-income countries.

The related source of data, mobile money, is also developing into another path for incredibly detailed insights, particularly in low- and middle-income countries (Centellegher et al., 2018). In countries where this service is available, people can use their mobile phone accounts as a form of bank account. Transactions between account holders get captured, allowing for a new glimpse into the micro-economic functioning of the communities of users of these services. During the COVID-19 pandemic, the use of mobile money both played a key role and evolved (Blumenstock, 2020; Chadha et al., 2020; McCaffrey, 2022).

Access to mobile data is an issue of contention due to a variety of factors and dimensions, and compelling arguments have been made to use it or to prevent its use. On the one hand, access to such data is clearly valuable for the insights it can provide. However, such access is controlled by the private sector (i.e., mobile phone providers), which has led to a very limited number of researchers and organizations being able to access this resource. From the standpoint of the impact such insights can bring, while their correct and careful analysis has the potential to generate positive change for the people being studied, it is also possible these findings could enhance the well-known disparities that the data revolution has helped foster, inadvertently or not. The balance between privacy and benefit is also an important topic, as clearly

mobile phone users can be uninformed participants in studies for which they have not given informed consent.

11.3.2 Earth Observation Data (Remote Sensing through Instruments such as Satellites)

Earth observation data derived from remote sensing instruments such as satellites (aka satellite data or remote sensing data) have been used over a number of decades—particularly luminosity—to provide high-resolution information about various local indicators, including population estimates (Welch, 1980), crop monitoring (Wójtowicz et al., 2016), and quantification of economic development (Jean et al., 2016), to name a few. The data are georeferenced over the surface of the earth, usually comprehensively and at high spatial resolution, although the temporal dimension is more limited, generally due to the constraints imposed by the satellite orbits. Data can consist of images from the visual spectrum but can also include other electromagnetic data, typically infrared. Given the vastness of the data, computational methods must be used to analyze them, including identification of structures and information. Image analysis and various statistical techniques including machine learning are often deployed on satellite data.

Earth observation data can support development work in particular where ground-based data (such as surveys or administrative records) are not available for areas or topics of interest (K. Anderson et al., 2017). Recently, these data have become more available and accessible due to the lowered cost of high-resolution imagery and the deployment of effective analytical techniques such as machine learning. Despite the promise they hold for the 2030 agenda (Andries et al., 2019), and the interest around earth observation data, the adoption of these data have been slow and uneven between countries (Kavvada et al., 2020).

Visual analysis of the environment can support projects with foci ranging from water, environment, and climate, to peace, and poverty reduction. Approaches using earth observation data to map levels of socio-economic development have gained popularity recently. Derived from nighttime satellite data and population density, the "Night Light Development Index" (Elvidge et al., 2012) offers a simple and globally available empirical measurement of human development. In Chapter 5 in this volume, Njuguna and McSharry use nightlight satellite data to estimate an index of multidimensional poverty in Rwanda. Using satellite images of night time light intensity

in a specific country or region, another study (Klomp, 2016) finds a correlation to income per capita and shows that as a result of natural disaster, the amount of light visible from space significantly decreases in the short term. An approach utilizing household survey data and Convolutional Neural Networks applied to high-resolution satellite images of cities has been used to generate detailed neighborhood-level maps which helped inform targeting for an anti-poverty program in Mozambique (Sohnesen et al., 2021), while Steele et al. (2017) have mapped poverty using mobile phone and satellite data. Satellite data are also increasingly used for research in environmental sustainability. Using satellite data on forest coverage along national borders, Crespo Cuaresma et al (2017) estimated the determinants of deforestation differences across countries and found that income per capita is the most important factor linked to lower forest coverage.

11.3.3 Sensor Data (In-situ Earth Observation and M2M Connectivity)

While earth observation data derived through remote sensing holds much promise for development, another source of geo-referenced data are in-situ sensors. Various stakeholders are increasingly experimenting with the adoption of connected sensors and machine-to-machine (M2M) connectivity in development initiatives and projects across many sectors. In the development arena, most applications utilize connected sensors which can detect a range of variables and physical information including (Hansen et al., 2013):

- position/presence/proximity
- temperature
- acoustic/sound/vibration
- flow
- leaks/levels
- acceleration/tilt
- motion/velocity/displacement
- humidity/moisture
- chemical/gas
- force/load/torque/strain/pressure
- electric/magnetic
- machine vision/optical/ambient light

A range of uses of Internet of Things (IoT) applications have emerged in contexts of development, where the interventions generally aim to improve

efficiency or effectiveness within service delivery, monitoring and evaluation, policy making, and research across various fields.[2] Within healthcare, sensors have been used to monitor the "cold chain" of vaccine delivery (R. Anderson et al., 2014; Lennon et al., 2017; Ross et al., 2020) and measure health outcomes of Ebola patients via a wearable "band-aid" (Steinhubl et al., 2015). Related to water and sanitation management, sensors have been deployed to monitor water quality and supply via integrated wireless networks (Zia et al., 2013), microbial fuel cells (Chouler & Di Lorenzo, 2015), pressure sensors (Apoorva et al., 2018), and smart handpumps (Papastylianou et al., 2014), a project which also led to the discovery of a method to estimate ground water depth (Colchester et al., 2017). Within agriculture, applications include micro-weather stations that observe field conditions and trigger indemnity payouts from an insurance scheme that protects farmers against weather-related extreme events (Wairimu et al., 2016; Altamirano & Beers, 2018), crop monitoring (Chaves, 2019), as well as evaluation of soil conditions (Zeitoun et al., 2021) and cooking patterns and adoption of more efficient cooking stoves (Wilson et al., 2015; Simons et al., 2018). IoT applications have also been introduced within financial services, education, and economic development projects (International Telecommunication Union, 2016).

11.3.4 Internet, Platforms, and Social Media

Another vast source of data can be derived from the Internet and, in particular, through digital platforms and social media channels. The rapid rate of digitalization and digitization has led to many interactions and transactions taking place online as well as given rise to new types of activities. Digital platforms collect a plethora of data on users and activities related to key issues in global development and can record real-time data on salient topics as well as offer a proxy for obtaining information that traditional data collection efforts may not be able to capture. For instance, data on skills and in particular digital skills is scarce, and analyzing user activity on platforms focused on collaborative programming may offer a window into computing skills in developing markets (Ojanperä et al., 2017, 2019). Job creation is a central challenge in many developing contexts, and platforms enable evaluation of job and labor

[2] The term Internet of Things describes the interconnection of the physical and virtual via an infrastructure comprised of devices that communicate with other devices (M2M) and enable the exchange of data between themselves. The contributing technologies tend to include sensors, RFID chips, nanotechnologies, and identification systems such as chips, cards, and SIMs (International Telecommunication Union, 2016).

market dynamics both in terms of digital labor as well as digital traces of the local labor markets (Scholz, 2013).

Search engines, another source of large digital data, are some of the most utilized websites on the Internet and search activity has been used to analyze a range of development themes such as the diffusion of e-commerce (Opesade, 2020), disease transmission dynamics (Majumder et al., 2016), and information seeking on HIV/AIDS (Okunoye et al., 2022). Social media channels such as X (formerly Twitter) and public information elements of Facebook can also shine a light on valuable questions for development. Among some examples for such use, the wealth index (Fatehkia et al., 2020) and the digital gender gap (Fatehkia et al., 2018) have been mapped using Facebook marketing data. There have also been active applications of these technologies such as Twitter vaccination campaigns (Alatas et al., 2019). There can be, however, cautionary tales. For example, Google Flu (a web service similar to Google Trends, which aggregated Google Search queries related to influenza), which showed considerable promise in offering a simple surveillance tool to track flu, generated a feedback mechanism by which Internet searches for terms associated with flu led to overestimation of severity of flu outbreaks (Lazer et al., 2014).

The size and origin of these data are quite varied. Some transactional data such as credit card records are highly sensitive and held in private hands. The issues that pertain to access of such data are quite similar to those for accessing mobile phone data. At the other end of the spectrum, some providers of online services offer open access to some of their information through so-called APIs (application programming interfaces). The comments made above regarding data size and techniques needed to analyze them apply regardless of openness or sensitivity.

11.3.5 Microfinance Data

Microfinance has emerged as an important financial tool to small enterprises and individuals in low-income settings. It includes a range of financial products such as small loans to support micro-enterprises as well as various other products to those outside of the traditional banking system, such as insurance or mortgages (Khavul, 2010). Where such data are accessible for research purposes, they can offer valuable insights about poverty reduction and financial inclusion.

While the data for these studies has come from different types of sources—e.g. proprietary data from micro-lending banks, or from large surveys—their

analysis has embraced the approaches associated with big data. Questions that have been addressed highlight the importance of social connections in the functioning of microfinance, and focus on how this resource is promoted by participants or non-participants (Banerjee et al., 2013), and the probability of repayment based on social network metrics (Sabin & Reed-Tsochas, 2020). Microfinance institutions collect large amounts of data about the financial activities of highly vulnerable populations.

While microfinance is driven with the goal of supporting venture creation, improving financial inclusion, and ultimately advancing development in settings of poverty, the impact of the various microfinancing programs has been mixed (Chen et al., 2017). This has turned research and discussion on microfinance towards the potential biases that come from using sensitive personal financial information, connecting the topic to others that address the detrimental impacts of using statistical techniques such as machine learning in predicting financial worthiness (Ashta & Herrmann, 2021). The main concern is exacerbation of patterns of inequity that, in principle, microfinance should be diminishing.

11.3.6 Surveys

Surveys have offered a longstanding and popular tool in development practice and research. Household surveys conducted face-to-face, postal surveys involving filling out a form, and field inventories or field mapping surveys recording the presence or absence of certain features of interest in the surveyed area have provided the main methods of gathering information from individuals and/or features of their social and natural environment. However, these methods of data collection are fairly expensive and not always practical, especially where short response times from a large or difficult-to-access population or geographic area are sought.

The increasing penetration of digital devices and wider connectivity have allowed traditional survey methods to be complemented with surveys that are deployed digitally to collect data remotely. Where the sample population can be accessed through email or directed to a URL, responses to fairly detailed surveys can be recorded conveniently, and shorter surveys can be quickly deployed through SMS and apps via mobile phones or devices connected to the Internet. These methods have been incorporated for instance by UNICEF's U-Report platform (2022) which collects data through SMS polls, alerts participants, and publishes the gathered data on various topics ranging from health to education and employment. Digitally enabled

survey methods might also assist in recording information in challenging contexts such as when asking about sensitive topics or when seeking to collect data from respondents with limited literacy. Audio Computer Assisted Self-Interviewing (ACASI) has been found to enable respondents to more readily share information on sensitive topics where they may limit their responses to an interlocutor in a face-to-face situation owing to embarrassment (Cullen, 2020) or where lack of reading or writing skills may limit the engagement with surveys conducted via text. The benefits of surveys deployed through digital avenues must be weighed against their limitations, including a possibly lower response rate in comparison to traditional surveys and the appropriateness of using a sampled population that is biased towards those with access to connected devices.

11.3.7 Traditional Administrative Data

Administrative data sources have been considered a gold standard in the literature for numerous areas of economics and other social sciences. Information such as official censuses, tax declarations, land records, death certificates, and other government-collected tabulations of population have provided considerable insight. As the data revolution has taken hold, there has been a simultaneous shift in the mindset and techniques associated with administrative data. Governments have begun to open pipelines to query administrative data (tax records, work records, etc.), making it possible for researchers to analyze data on the full population. The movement for open development data has further increased the accessibility of many administrative datasets. The analytical approach to those data has changed as well. Traditional methodologies such as regression analysis and hypothesis testing have been joined by new methods such as network science, machine learning, GIS, and a variety of other approaches—engendering a new view of these "old" resources.

A key variable in the use of these data in the present, and especially into the future, is practical: how easily accessible can such data be? Quality and ease (with appropriate safeguards) can maximize the utility of this traditional resource. Over the last few decades, initiatives and programs related to open data have made an increasing amount of traditional data sources accessible (Open Government Partnership, 2023). Two relatively recent uses of administrative data in the context of development can be found for intergenerational mobility in India (Asher et al., 2020) and education dropout in Guatemala and Honduras (Adelman et al., 2018). The insightful review by Ruggles (2013) remains current. In this review, a summary of the many

disparate sources of large microdata, efforts at integrating them, and the new models of access to it are discussed, including comments on how these sources are also being integrated with big data, among others.

11.4 Summary

Sources of big data are not static. For example, richer forms of mobility data have only emerged in recent years with the rise of smartphones. More recently, data deriving from the Internet of Things is only starting to be utilized for the purposes of global development (International Telecommunication Union, 2016). Other changes will come in the future.

The most important lessons about this chapter—in view of this constant change—rest in the diverse nature of big data, their power, and their limitations. Big data are not data of everything or everyone. These data capture parts of a situation, sometimes most people, sometimes some people. The data sampling can be biased and incomplete and collected for purposes distinct from its envisioned uses in development. Usually, no controls are associated with the data. For the most part, one can think of big data as coming from so-called natural experiments, and they should be analyzed with this in mind (Einav & Levin, 2014).

In closing—we should always look at big data with a critical eye.

References

Achard, F., Eva, H. D., Stibig, H.-J., Mayaux, P., Gallego, J., Richards, T., & Jean-Paul Malingreau. (2002). Determination of deforestation rates of the world's humid tropical forests. *Science* (New York), *297*(5583), 999–1002. https://doi.org/10.1126/science.1070656

Adelman, M., Haimovich, F., Ham, A., & Vazquez, E. (2018). Predicting school dropout with administrative data: New evidence from Guatemala and Honduras. *Education Economics*, *26*(4), 356–372. https://doi.org/10.1080/09645292.2018.1433127

Alatas, V., Chandrasekhar, A. G., Mobius, M., Olken, B. A., & Paladines, C. (2019). *When Celebrities Speak: A Nationwide Twitter Experiment Promoting Vaccinations in Indonesia* (Working Paper 25589; CESIfo Working Paper). National Bureau of Economic Research.

Altamirano, M. A., & Beers, C. P. van. (2018). Frugal innovations in technological and institutional infrastructure: Impact of mobile phone technology on productivity, public service provision and inclusiveness. *The European Journal*

of Development Research, *30*(1), 84–107. https://doi.org/10.1057/s41287-017-0115-2

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., & Friedl, L. (2017). Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-Spatial Information Science*, *20*(2), 77–96. https://doi.org/10.1080/10095020.2017.1333230

Anderson, R., Perrier, T., Pervaiz, F., Sisouveth, N., Kumar, B., Phongphila, S., Rahman, A., Dhiman, R., & Newland, S. (2014). Supporting immunization programs with improved vaccine cold chain information systems. *IEEE Global Humanitarian Technology Conference (GHTC 2014)*, 215–222. https://doi.org/10.1109/GHTC.2014.6970284

Andries, A., Morse, S., Murphy, R., Lynch, J., Woolliams, E., & Fonweban, J. (2019). Translation of Earth observation data into sustainable development indicators: An analytical framework. *Sustainable Development*, *27*(3), 366–376. https://doi.org/10.1002/sd.1908

Ansón, J., & Helble, M. (2013). A Gravity Model of International Postal Exchanges. In M. A. Crew & P. R. Kleindorfer (Eds.), *Reforming the Postal Sector in the Face of Electronic Competition* (pp. 36–47). Edward Elgar Publishing.

Apoorva, R., Biswas, D., & Srinivasan, V. (2018). Do household surveys estimate tap water use accurately? Evidence from pressure-sensor based estimates in Coimbatore, India. *Journal of Water, Sanitation and Hygiene for Development*, *8*(2), 278–289. https://doi.org/10.2166/washdev.2018.127

Aron, J. (2018). Mobile money and the economy: A review of the evidence. *The World Bank Research Observer*, *33*(2), 135–188. https://doi.org/10.1093/wbro/lky001

Asher, S., Novosad, P., & Rafkin, C. (2020). Intergenerational mobility in India: New methods and estimates across time, space, and communities. *Unpublished Manuscript. Available at*: https://Paulnovosad.Com/Pdf/Anr-India-Mobility.Pdf.

Ashta, A., & Herrmann, H. (2021). Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance. *Strategic Change*, *30*(3), 211–222. https://doi.org/10.1002/jsc.2404

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *Science* (New York), *341*(6144), 1236498.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code* (1st edition). Polity.

Björkegren, D., & Grissen, D. (2019). Behavior revealed in mobile phone usage predicts credit repayment. *The World Bank Economic Review*, *34*(3), 618–634. https://doi.org/10.1093/wber/lhz006

Blumenstock, J. (2020). Machine learning can help get COVID-19 aid to those who need it most. *Nature*. https://doi.org/10.1038/d41586-020-01393-7

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Bogomolov, A., Lepri, B., Larcher, R., Antonelli, F., Pianesi, F., & Pentland, A. (2016). Energy consumption prediction using people dynamics derived from cellular network data. *EPJ Data Science*, *5*(1), 13. https://doi.org/10.1140/epjds/s13688-016-0075-3

Bratley, K., & Ghoneim, E. (2018). Modeling urban encroachment on the agricultural land of the Eastern Nile Delta using remote sensing and a GIS-based Markov Chain Model. *Land*, *7*(4), Article 4. https://doi.org/10.3390/land7040114

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLOS Medicine*, *5*(7), 1–6. https://doi.org/10.1371/journal.pmed.0050151

Centellegher, S., Miritello, G., Villatoro, D., Parameshwar, D., Lepri, B., & Oliver, N. (2018). Mobile money: Understanding and predicting its adoption and use in a developing economy. *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies*., *2*(4). https://doi.org/10.1145/3287035

Cesare, N., Lee, H., McCormick, T., Spiro, E., & Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography*, *55*(5), 1979–1999. https://doi.org/10.1007/s13524-018-0715-2

Chadha, S., Kipkemboi, K., & Muthiora, B. (2020, July 16). Tracking mobile money regulatory responses to COVID-19. *Mobile for Development*. https://www.gsma.com/mobilefordevelopment/programme/mobile-money/gsma-mobile-money-regulatory-response-to-covid-19-tracker-and-analysis/

Chaves, M. (2019). Recent applications of the MODIS sensor for soybean crop monitoring and deforestation detection in Mato Grosso, Brazil. *CAB Reviews Perspectives in Agriculture Veterinary Science Nutrition and Natural Resources*, *14*. https://doi.org/10.1079/PAVSNNR201914007

Chen, J., Chang, A. Y., & Bruton, G. D. (2017). Microfinance: Where are we today and where should the research go in the future? *International Small Business Journal*, *35*(7), 793–802. https://doi.org/10.1177/0266242617717380

Chouler, J., & Di Lorenzo, M. (2015). Water quality monitoring in developing countries; Can microbial fuel cells be the answer? *Biosensors*, *5*(3), Article 3, 450–470. https://doi.org/10.3390/bios5030450

Colchester, F. E., Marais, H. G., Thomson, P., Hope, R., & Clifton, D. A. (2017). Accidental infrastructure for groundwater monitoring in Africa. *Environmental Modelling & Software*, *91*, 241–250.

Crespo Cuaresma, J., Danylo, O., Fritz, S., McCallum, I., Obersteiner, M., See, L., & Walsh, B. (2017). Economic development and forest cover: Evidence from satellite data. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/srep40678

Cullen, C. (2020). *Method Matters: Underreporting of Intimate Partner Violence in Nigeria and Rwanda* (SSRN Scholarly Paper 3624515). Social Science Research Network. https://papers.ssrn.com/abstract=3624515

Dalton, C., & Thatcher, J. (2014). *What Does A Critical Data Studies Look Like, And Why Do We Care?* https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J.-M., Krings, G., Gutierrez, T., Blondel, V. D., & Luengo-Oroz, M. A. (2014). *Estimating Food Consumption and Poverty Indices with Mobile Phone Data*. arXiv. https://doi.org/10.48550/ARXIV.1412.2595

de Montjoye, Y. A., Smoreda, Z., Trinquart, R., Ziemlicki, C., & Blondel, V. D. (2014). *D4D-Senegal: The Second Mobile Phone Data for Development Challenge*. https://arxiv.org/abs/1407.4885

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

D'ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Eagle, N., Pentland, A. (Sandy), & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278. https://doi.org/10.1073/pnas.0900282106

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089. https://doi.org/10.1126/science.1243089

Elvidge, C. D., Baugh, K. E., Anderson, S. J., Sutton, P. C., & Ghosh, T. (2012). The Night Light Development Index (NLDI): A spatially explicit measure of human development from satellite data. *Social Geography*, *7*(1), 23–35. https://doi.org/10.5194/sg-7-23-2012

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Fatehkia, M., Kashyap, R., & Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Development*, *107*, 189–209. https://doi.org/10.1016/j.worlddev.2018.03.007

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., & Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, *9*(1), 22. https://doi.org/10.1140/epjds/s13688-020-00235-w

The Flowminder Foundation. (2023). *Flowminder—Innovative Solutions fr Mobile Data Analytics*. https://www.flowminder.org

Graham, M., & Dittus, M. (2022). *Geographies of Digital Exclusion: Data and Inequality*. Pluto Press. https://cir.nii.ac.jp/crid/1130293004051685790

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & J. R. G. Townshend. (2013). High-resolution

global maps of 21st-century forest cover change. *Science* (New York), *342*(6160), 850–853. https://doi.org/10.1126/science.1244693

Hristova, D., Rutherford, A., Anson, J., Luengo-Oroz, M., & Mascolo, C. (2016). The international postal network and other global flows as proxies for national wellbeing. *PLOS ONE*, *11*(6), 1–19. https://doi.org/10.1371/journal.pone.0155976

International Telecommunication Union. (2016). *Harnessing the Internet of Things for Global Development*. https://www.itu.int/en/action/broadband/Documents/Harnessing-IoT-Global-Development.pdf

Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., Pentland, A. (Sandy), & de Montjoye, Y.-A. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Science*, *6*, 1–21.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Kavvada, A., Metternicht, G., Kerblat, F., Mudau, N., Haldorson, M., Laldaparsad, S., Friedl, L., Held, A., & Chuvieco, E. (2020). Towards delivering on the Sustainable Development Goals using Earth observations. *Remote Sensing of Environment*, *247*, 111930. https://doi.org/10.1016/j.rse.2020.111930

Khavul, S. (2010). Microfinance: Creating opportunities for the poor? *Academy of Management Perspectives*, *24*(3), 58–72. https://doi.org/10.5465/AMP.2010.52842951

Kibanov, M., Stumme, G., Amin, I., & Lee, J. G. (2017). Mining social media to inform peatland fire and haze disaster management. *Social Network Analysis and Mining*, *7*(1), 30. https://doi.org/10.1007/s13278-017-0446-1

Kim, J., Cha, M., & Lee, J. G. (2017). Nowcasting commodity prices using social media. *PeerJ Computer Science*, *3*, e126.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Klomp, J. (2016). Economic development and natural disasters: A satellite data analysis. *Global Environmental Change*, *36*, 67–88. https://doi.org/10.1016/j.gloenvcha.2015.11.001

Koehler, J., Thomson, P., & Hope, R. (2015). Pump-priming payments for sustainable water services in rural Africa. *World Development*, *74*, 397–411. https://doi.org/10.1016/j.worlddev.2015.05.020

Kurekin, A., Loveday, B., Clements, O., Quartly, G., Miller, P., Wiafe, G., & Adu Agyekum, K. (2019). Operational monitoring of illegal fishing in Ghana through exploitation of satellite earth observation and AIS data. *Remote Sensing*, *11*(3), 293. https://doi.org/10.3390/rs11030293

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science* (New York), *343*, 1203–1205.

Lennon, P., Atuhaire, B., Yavari, S., Sampath, V., Mvundura, M., Ramanathan, N., & Robertson, J. (2017). Root cause analysis underscores the importance of understanding, addressing, and communicating cold chain equipment failures to improve equipment performance. *Vaccine*, *35*(17), 2198–2202. https://doi.org/10.1016/j.vaccine.2016.09.068

Liang, L., Shrestha, R., Ghosh, S., & Webb, P. (2020). Using mobile phone data helps estimate community-level food insecurity: Findings from a multi-year panel study in Nepal. *PLOS ONE*, *15*(11), 1–16. https://doi.org/10.1371/journal.pone.0241791

Majumder, M. S., Santillana, M., Mekaru, S. R., McGinnis, D. P., Khan, K., & Brownstein, J. S. (2016). Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015–2016 Colombian Zika virus disease outbreak. *JMIR Public Health and Surveillance*, *2*(1), e5814. https://doi.org/10.2196/publichealth.5814

Marquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Martinez, E. A., Rubio, M. H., Martinez, R. M., Arias, J. M., Patane, D., Zerbe, A., Kirkpatrick, R., Luengo-Oroz, M., & Zerbe, A. (2016). *Measuring Economic Resilience to Natural Disasters with Big Economic Transaction Data*. arXiv. https://doi.org/10.48550/ARXIV.1609.09340

McCaffrey, M. (2022, June 30). *Untold Stories from 15 Years of Mobile Money Data*. UNCDF - Impact Capital for Development. https://www.uncdf.org/article/7823/did-the-covid-19-pandemic-accelerate-growth-of-mobile-money

Menon, R., Biswas, A., Armin, S., Quinn, J., & Niesler, T. (2018). Automatic Speech Recognition for Humanitarian Applications in Somali. *Proceedings of SLTU*. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), Gurugram, India.

Menon, R., Saeb, A., Cameron, H., Kibira, W., Quinn, J., & Niesler, T. (2017). Radio-browsing for developmental monitoring in Uganda. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5795–5799. https://doi.org/10.1109/ICASSP.2017.7953267

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://nyupress.org/9781479837243/algorithms-of-oppression

Ojanperä, S., Graham, M., Straumann, R., Sabbata, S. D., & Zook, M. (2017). Engagement in the knowledge economy: Regional patterns of content creation

with a focus on Sub-Saharan Africa. *Information Technologies & International Development*, *13*(0), 19.

Ojanperä, S., Graham, M., & Zook, M. (2019). The Digital Knowledge Economy Index: Mapping content production. *The Journal of Development Studies*, *55*(12), 2626–2643. https://doi.org/10.1080/00220388.2018.1554208

Okunoye, B., Ning, S., & Jemielniak, D. (2022). Searching for HIV and AIDS Health information in South Africa, 2004–2019: Analysis of Google and Wikipedia search trends. *JMIR Formative Research*, *6*(3), e29819. https://doi.org/10.2196/29819

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Open Government Partnership. (2023). *The Future of Open Government*. Open Government Partnership. https://www.opengovpartnership.org/

Opesade, A. O. (2020). Discovering patterns in electronic commerce diffusion in Nigeria using Google Trends web data. *Journal of Information Science, Systems and Technology*, *4*(1), 1–29.

Papastylianou, T., Behar, J., Guazzi, A., Jorge, J., Laranjeira, S., Maraci, M. A., Clifford, G. D., Hope, R. A., & Thomson, P. (2014). *Smart Handpumps: Technical Aspects of a One-year Field Trial in Rural Kenya*. 9–9. https://doi.org/10.1049/cp.2014.0769

Patel, A., & Soni, J. B. (2022). Bhoomi Prahari—e-governance tool for monitoring encroachment on government land using mobile and GIS technology. *International Journal of Information Technology*, *14*(3), 1565–1573. https://doi.org/10.1007/s41870-021-00674-2

Pérez, B., Castellanos, C., & Correal, D. (2018). Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study. In A.D. Orjuela-Cañón, J.C. Figueroa-García, & J.D. Arias-Londoño (Eds), *Applications of Computational Intelligence* (pp. 111–125). Springer International Publishing. https://doi.org/10.1007/978-3-030-03023-0_10

Ross, J. C., Saidu, Y., Nzuobontane, D., Voukings, M. Z., & Embrey, S. R. (2020). Application of the remaining vaccine vial monitor life calculation to field temperature monitoring data to improve visibility into cold chain equipment performance. *Vaccine*, *38*(48), 7683–7687. https://doi.org/10.1016/j.vaccine.2020.09.078

Ruggles, S. (2013). Big microdata for population research. *Demography*, *51*(1), 287–297. https://doi.org/10.1007/s13524-013-0240-2

Sabin, N., & Reed-Tsochas, F. (2020). Able but unwilling to enforce: Cooperative dilemmas in group lending. *American Journal of Sociology*, *125*(6), 1602–1667. https://doi.org/10.1086/709105

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., & Vespignani, A.

(2012). Digital epidemiology. *PLOS Computational Biology*, *8*(7), 1–3. https://doi.org/10.1371/journal.pcbi.1002616

Scholz, T. (2013). *Digital Labor: The Internet as Playground and Factory*. Routledge.

Shupler, M., O'Keefe, M., Puzzolo, E., Nix, E., Anderson de Cuevas, R., Mwitari, J., Gohole, A., Sang, E., Čukić, I., Menya, D., & Pope, D. (2021). Pay-as-you-go liquefied petroleum gas supports sustainable clean cooking in Kenyan informal urban settlement during COVID-19 lockdown. *Applied Energy*, *292*, 116769. https://doi.org/10.1016/j.apenergy.2021.116769

Simons, A. M., Beltramo, T., Blalock, G., & Levine, D. I. (2018). Sensor data to measure Hawthorne effects in cookstove evaluation. *Data in Brief*, *18*, 1334–1339. https://doi.org/10.1016/j.dib.2018.04.021

Sohnesen, T. P., Fisker, P., & Malmgren-Hansen, D. (2022). Using satellite data to guide urban poverty reduction. *Review of Income and Wealth*, *68*, S282–S294. https://doi.org/10.1111/roiw.12552

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J., & Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, *14*(127), Article 20160690.

Steinhubl, S. R., Marriott, M. P., & Wegerich, S. W. (2015). Remote sensing of vital signs: A wearable, wireless "band-aid" sensor with personalized analytics for improved Ebola patient care and worker safety. *Global Health: Science and Practice*, *3*(3), 516–519. https://doi.org/10.9745/GHSP-D-15-00189

Toole, J. L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M. C., & Lazer, D. (2015). Tracking employment shocks using mobile phone data. *Journal of the Royal Society Interface*, *12*(107), 20150185. https://doi.org/10.1098/rsif.2015.0185

UN Global Pulse. (2017). *Using Big Data Analytics for Improved Public Transport*. UN Global Pulse. https://pulselabjakarta.org/assets/uploadworks/2018-08-02-13-21-05.pdf

UNICEF. (2022). *U-Report*. https://www.unicef.org/innovation/U-Report

United Nations Development Group. (2017). *Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda*. United Nations Development Group. undg.org/document/data-privacy-ethics-and-protection-guidance-note-on-big-data-for-achievement-of-the-2030-agenda/

Wairimu, E., Obare, G., & Odendo, M. (2016). Factors affecting weather index-based crop insurance in Laikipia County, Kenya. *Journal of Agricultural Extension and Rural Development*, *8*(7), 111–121. https://doi.org/10.5897/JAERD2016.0769

Welch, R. (1980). Monitoring urban population and energy utilization patterns from satellite data. *Remote Sensing of Environment*, *9*(1), 1–9.

Wilson, D. L., Adam, M. I., Abbas, O., Coyle, J., Kirk, A., Rosa, J., & Gadgil, A. J. (2015). Comparing Cookstove Usage Measured with Sensors Versus Cell Phone-Based Surveys in Darfur, Sudan. In S. Hostettler, E. Hazboun, & J.-C. Bolay (Eds), *Technologies for Development* (pp. 211–221). Springer International Publishing. https://doi.org/10.1007/978-3-319-16247-8_20

Wójtowicz, M., Wójtowicz, A., & Piekarczyk, J. (2016). Application of remote sensing methods in agriculture. *Communications in Biometry and Crop Science*, *11*(1), 31–50.

Zeitoun, R., Chaudhry, H., Vasava, H. B., & Biswas, A. (2021). Global Development in Soil Science Research: Agriculture Sensors and Technologies. In A. Rakshit, S. K. Singh, P. C. Abhilash, & A. Biswas (Eds), *Soil Science: Fundamentals to Recent Advances* (pp. 599–612). Springer. https://doi.org/10.1007/978-981-16-0917-6_29

Zeug, H., Zeug, G., Bielski, C., Solano-Hermosilla, G., & M'Barek, R. (2017). *Innovative food price collection in developing countries: Focus on crowdsourcing in Africa. 28247*. http://publications.europa.eu/publication/manifestation_identifier/PUB_LBNA28247ENE

Zhang, G., & He, B.-J. (2021). Research on the Global Green Market based on Big Data. In D. La Rosa & R. Privitera (Eds), *Innovation in Urban and Regional Planning* (pp. 139–148). Springer International Publishing.

Zia, H., Harris, N. R., Merrett, G. V., Rivers, M., & Coles, N. (2013). The impact of agricultural activities on water quality: A case for collaborative catchment-scale management using integrated wireless sensor networks. *Computers and Electronics in Agriculture*, *96*, 126–138. https://doi.org/10.1016/j.compag.2013.05.001

12

Methods for the Analysis of Big Data in Development

Eduardo López and Sanna Ojanperä

12.1 Introduction

There is no simple way to attempt a summary of data analysis techniques that can do the topic justice. Limiting the summary to the boundaries of big data in global development does not lead to any great advantages either, except for the fact that slow methods of data analysis become less critical to discuss, as they cannot handle the data volume.

Therefore, in an attempt to provide the simplest usable description of the methodological aspects of the use of big data for human development, in this chapter we modestly present some basic ideas about the context in which these methods live (the world of probabilities), a discussion of the evolution of data from "small" to "big," how this has changed methodologies and ways of thinking, and finally enumerate and briefly describe some approaches that have been used in development work that incorporate big data.

Note that many sources of big data originate in specialized settings (e.g., satellite images or GPS signals). The details of the methods applied to generate data in these settings are not covered in this chapter. Instead, we focus solely on the methods that are applied to data that result from the data creation process, which usually capture aspects of human behavior. Effectively, our discussion will be about a variety of general statistical methods which are common to many big data and development projects.

12.2 Big Data and the Traditional Toolboxes in Development Cooperation and Research

Before we review and discuss the methodological approaches relevant for big data in development, it is also useful to reflect on traditional methodological

Eduardo López and Sanna Ojanperä, *Methods for the Analysis of Big Data in Development*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press.
© Eduardo López and Sanna Ojanperä (2025). DOI: 10.1093/oso/9780198879145.003.0013

approaches. Given the range of actors and paradigms of development, there is considerable diversity in the scope of the professional roles within development work. While economists, lawyers, or spatial analysts are often found within development work, many working in this area have undertaken training specific to the field such as research, action research, practice, or policy related to development. Whereas the professional communities and training pathways for more narrowly defined professional roles that rely heavily on quantitative analysis such as economist or spatial analyst are likely to be more open to working with large datasets or adopting new analytical methods, the vast majority of the development research, action research, practice, and policy oriented professional communities and training pathways continue to rely on more traditional quantitative tools.

The skills of the professionals just mentioned are often acquired during formal education, and most of the training directed toward equipping individuals with capacities for development work tends to take place within social science faculties. This disciplinary focus is reflected in the methodological toolkit, which mainly draws from social research methodologies and statistical approaches in particular, including applied policy analysis, micro- and macroeconomic analysis, econometrics, and geographic information systems (GIS). While individuals often choose between building qualitative or quantitative research skills, the potential of the complementarity of the two approaches is worth highlighting, as such "mixed methods" research approaches offer much potential for development work (Harris, 2022). However, as this edited volume focuses on mainly quantitative approaches, we leave aside further discussions on the qualitative and mixed methods approaches. The methodological toolbox of those development professionals with quantitative training variably includes basic statistical concepts and descriptive statistical methods (such as sampling methods, comparison of means, analysis of contingency tables, correlation, and regression as well as an understanding of central concepts such as game theory, public economics, welfare theorems, and opinion dynamics models) and statistical methods aiming to address causal inference (such as longitudinal and panel data analysis, project and program evaluation, experimental design and analysis, survey design and analysis, principal components analysis, factor analysis, multilevel models, and other econometric tools). Often, significant time is dedicated to building skills for the development of research designs and data collection, as project and program evaluation exercises frequently rely on the deployment of household surveys or experiments.

The majority of these methods and concepts are well-established and reflect the statistical nature of the data analysis problem inherent to development. This is fortunate because, in fact, one advantage of big data is that they enable the use of many of these techniques to develop insight. However, the techniques are often taught using software tools such as SPSS or Stata that have a limited capacity for computation and tend to lack the robustness necessary to handle large datasets. Specialized bridging training is required to apply these methods with large datasets using software packages such as R, scripting languages like Python, or mid-level programming languages like C (or its derivatives). This approach of going back to basic programming lends robustness to the data analysis and is becoming standard practice in the space of data-driven business analytics. In Python, in particular, many open-source libraries have been recently developed to handle scientific computing but owing to the openness and decentralized nature of these tools, do not follow as uniform a functional logic as that of proprietary software. Relatedly, the application of these methods in the context of secondary big data sources requires a thorough understanding of the possibilities and limitations of big data—a topic which is often absent in the training, professional discussions, or fora of those working in development (Bennett et al., 2022). Another aspect of this transition to bigger datasets is the not-too-infrequent need for the creation of new methods to handle the large data volume, a task that can only be performed by those whose training goes beyond the mere application of packages and stretches into a conceptual understanding of the underlying quantitative nature of the problem and the related data.

There are many benefits to be reaped from the use of big data in contexts of development work, in both research-related as well as practice-focused roles. Where relevant for research efforts, big secondary data and related methods may offer an important source of evidence for evaluating the impacts of projects and programs and reduce the significant time and monetary cost of designing and running surveys. In the context of development practice, visualizing large datasets can help communication and facilitate discussions around varied topics. The next sections of the chapter will outline methods and applications of big data in development to help offer insight in this area.

12.3 The Common Probabilistic Setting of Data

While not all data (including big data) can be understood as probabilistic, much of them can be. This is particularly true when those data relate to human behavior. Two people faced with the same set of circumstances, and

even characterized by virtually indistinguishable personal features, can and often do react differently. This means that most of the data to be analyzed have to be looked at from the perspective of probability and statistics.

Let us discuss the basics of any probabilistic settings. In a defined circumstance or context (for example subsistence farmers making decisions about whether or not to use fertilizers on a given crop) there is a group of possible events that can happen (the farmers may decide to use or forego fertilizers); in technical terms, the things that can happen make up the so-called *sample space*. Each of the things that can happen has a number between 0 and 1 attached to it, called its probability (e.g., 0.3 for using fertilizer). When all these numbers are added up, they sum to 1, which is the probabilistic way of saying that all possibilities add up to 100% (e.g., since 0.3 quantifies the chances of using fertilizers, 0.7 are the chances of not using them).

A general way to think about probabilistic settings is to imagine a situation in which a group of people with a collection of $n - 1$ known features (say age, gender, salary, occupation) that have the same values (say, 20 years of age, male, making $20,000 a year, all car mechanics) can still lead to an n^{th} feature that can take more than one possible value without certainty (e.g., they vote for a different political party). Furthermore, this can be generalized so that there may be $n - r$ common feature values (people with the same age, salary, education) and r features with distinct outcomes (some have children and others do not, some donate to charities, but others do not, etc.). One of the known features can be time, which in the framework of econometrics would be called panel data and in several natural science disciplines and engineering would be called a time series.

A key part of understanding probabilistic settings is that the most one can know about them is the collection of probabilities about each and every possible outcome. In other words, we can know the sample space completely and know all the probabilities of each of its elements. This understanding is at the core of much of the quantitative analysis of socioeconomic systems. Both statistical characterization and time series forecasting can be treated with the same conceptual basis, although the techniques change depending on which of these two contexts apply.

Once the sample space and the probabilities of a problem are known, one can know everything else that is *probabilistically knowable*. This means that given enough information, we can determine sample space and probabilities of the elements of that sample space, and from that we can make any statistical estimate that we need. This suggests that much can be known provided there is sufficient data.

Unfortunately, in many problems in social science in general, and in development, specifically, one is often "feeling in the dark." This is because we hardly ever have enough information. In many real situations, there are great challenges to knowing all (or indeed any) parts of this probabilistic setting. For example, it is typical to know only parts of the sample space, and because of that, there are possible, small-probability outcomes that one ends up missing, leading to underestimation of these missing outcomes and overestimation of the other probabilities. This provides a false sense of confidence in one's ability to estimate. For example, subsistence farmers may delay the use of the fertilizer, a possibility one may not have known and thus failed to consider.

Even if the sample space is known, another typical problem is the inability to properly estimate probabilities. In practice, probabilities are estimated by seeing the same situation many times and being able to capture all the outcomes often enough that one knows the correct relative proportions with which outcomes occur. In reality, we hardly ever see anything enough times. Big data sometimes offer relief from this problem due to their size, but in effect, one may also end up being confused. That is because the data may not (as is often the case) have captured all of the relevant features that describe the event, or maybe because the number of combinations of things that can happen is so vast that even with all features captured, one does not have enough observations to estimate properly. Only in the cases where all possible relevant features are captured with enough sampling, can the amount of data (big or otherwise) be enough.

The final and most difficult problem stems from not even picking the right context. This is far less trivial than appears. We have illustrated the sample space concept from the collection of decisions a subsistence farmer may take about a crop (using fertilizers or not). But what if that choice of items that gives us enough information for the decision has not been collected completely? For instance, it is plausible that one should also keep track of the weather in order to understand the problem correctly. Or maybe adding weather is still not enough because it is necessary to also consider prices of crop-related products used by the farmer. These additional elements of information are called by different names, and traditionally have been referred to as *variables*, but in recent times following the growing popularity of increasingly computationally intensive methodological approaches, the terms *features* or *attributes* have become very common. As becomes obvious fairly quickly, when one starts to consider more details (attributes), more questions arise. In the end, the approach taken in most circumstances is to draw a conceptual boundary around the situation, tracking some variables

and ignoring others. In the most prudent approach, the boundary is chosen either by applying principles that dictate those boundaries, or by testing the approach so that it minimizes the effects of the untracked information. In other cases, however, such careful control is not exercised or may not even be possible (data cannot be collected, laws prevent such collection or use, or decisions about how to study the probabilistic setting are made without consideration of proper control). And if one were able to gather all the data possible, we would very quickly return to the poor sampling situation, where the possibilities are so vast that there are never enough samples.

We should note that the description above is based on what is called a discrete (and finite) context. While the same concepts apply, sample spaces do not have to be discrete. For example, the important question of fertilizer use may not be whether to use it or not, but rather how much. Immediately this results in the size of the sample space becoming infinite as the amount of fertilizer can be prescribed down to any level of precision (milligrams, micrograms, nanograms, etc.). Of course, in practice, the problems are usually converted back to some discrete form (e.g., use between 10 and 20 grams of fertilizer, or never more than 30, etc.).

As this discussion illustrates, the reality of data analysis is that it must estimate incompletely defined probabilities with limited information. But there is hope in that those estimates even under challenging conditions do often yield actionable information. The ingenuity of the methods, accompanied by the fact that many situations, even if influenced by many factors, are typically controlled by only a few, make it so that methods do work, and most of the art and expertise in data analysis is associated with picking the relevant data to be analyzed through the right methods.

So, as we delve into the methods, keep in mind some guiding questions regarding any problem being studied with a given technique: What is the correct (or most reliable) context one should focus on? What are the possible outcomes in that context? What are their probabilities? How are all these pieces of information being generated (both the data and the estimations)? What is being done with this information?

12.4 The Changing Situation: From "Small" Data to "Big Data" and the Data Revolution

The development of large data sources has generated an important dichotomy in the choice of approaches. Traditional methods to analyze data relevant for development have relied on statistical techniques deployed in

disciplines such as economics, geography, and a variety of social sciences. Many of these techniques are useful for data sets with small sizes, and the methods are effective at finding information from weak signals (i.e., uncovering interesting trends, patterns, or causal relationships in the data).

As large data have come into the picture, the applicability of traditional techniques has required rethinking. For example, summary statistics, regression, and correlation analysis which, by their nature, are computationally efficient, have continued to be useful in this new data landscape. Other methodologies that require intensive computation on small data sets have become less useful as the size of new data prevents their application.

The strength of signal due to the vast amounts of data in the Big Data[1] setting eases the situation: simple techniques can yield considerable information without the need to appeal to demanding and subtle statistical techniques. On the other hand, the mere existence of larger data cannot guarantee that signal is easy to find, if it is even present. Furthermore, large data can elicit more ambitious and detailed questions which may once again depend on weaker signals and new, subtle techniques.

Therefore, as a natural response to this, a vast array of new techniques have emerged which seek to be *scalable* (meaning that they can deal with growing sizes of data more easily and efficiently) and automated (needing less human intervention), among other traits (Smyth, 2000). An important component of the challenge is that, with data that contains many attributes, the number of ways in which it can be analyzed increases rapidly. This has fueled the emergence of sophisticated fields of methods to extract information such as Machine Learning (ML), which attempt to uncover information through "learning" from the data (Poggio & Smale, 2003). Along with traditional techniques, these new collections of methods are able to seek signal from large data sets with various levels of human intervention (from supervised to unsupervised) and offer insights into many situations. Cutting the human out of the loop has advantages and disadvantages: humans slow down the data analysis process which means that very large data, especially with many combinations of features, can become effectively impossible to fully explore, leading to only a subset of possible interesting questions being studied. On the other hand, when machine algorithms explore the data in an unsupervised manner, humans no longer slow the process down, but the meanings extracted from the data by algorithms can be, in many cases, questionable or

[1] Here, we capitalize "Big Data" to discuss it as a concept as defined in Chapter 4 by Letouzé.

even useless. Because of these tradeoffs, it is common to see partly or highly supervised application of machine learning methods. Much of what is at the core of this performance tradeoff is the possibility of uncovering causal mechanisms out of the data. An important criticism of big data is that, because they are not captured with the appropriate controls designed into academic studies (Einav & Levin, 2014; Kitchin, 2014), they may be incapable of uncovering causality.

This very question of causality has influenced the nature of questions asked of the data. While traditional approaches have been more focused on understanding *why* a situation takes place, many large-data approaches are content to leave causal mechanisms aside and focus on the *probability* that something occurs (now or in the future). This way of thinking is encapsulated in the provocative statements about "the end of theory" (R. Anderson et al., 2014) proffered by various people in both industry and academia. Although much of this idea has been justly criticized (Kitchin, 2014), we should acknowledge that there is value in addressing concrete questions without thought to causality because, to tackle on-the-ground situations, we do not always need to know their origin: the problems faced by populations in contexts of poverty, lack of opportunity, and other challenges should not wait for an explanation before something tangible is done about them, even if it is palliative. While the need to continue the scientific inquiry into the root causes of underdevelopment and lack of opportunity has a guaranteed place in data analysis, this recent engineering-like approach about how to use data (aimed at application rather than theory) offers a new pragmatic and valuable perspective.

One more consideration, touched on in other chapters of this volume, is the change in mentality that Big Data has brought about, the so-called data revolution. The choice of questions that researchers and practitioners are approaching has evolved, even when exploring older traditional data sets. For example, there is greater interest in dual micro–macro perspectives that attempt to address phenomena at the system and individual levels than ever before, whereas traditionally aggregation was the norm.

Taking all these considerations into account, it would be fair to say that the present state of affairs regarding methodological approaches to the study of data relevant to development is highly heterogeneous. If we were to try to classify the various methods, we would want to consider where they lie along several dimensions: concerned with causality or pure prediction, new or traditional, and scalable in size or poorly scaling. However, as indicated before, we set our sights on a more modest classification based on usage alone.

12.5 Methods

The next step is to go from big data to actionable information. For those entering the field, the landscape can be overwhelming, largely due to the sheer volume of methods available. This abundance, in turn, reflects the variety of empirical settings that generate data to be analyzed.

However, a review of many projects analyzing big data in development both in academic and practitioner contexts reveals that there are common analytical tools that are used extremely frequently. Unsurprisingly, they are also some of the most classic and common methodologies, including summary statistics, time series analysis, correlations, and regressions, to name a few. All of these methodologies satisfy the scalability criterion mentioned above. With regards to new approaches, the most prominent examples are Machine Learning (Mitchell, 1997), Natural Language Processing (Eisenstein, 2019), and Network Science (Newman, 2018). Another pattern to notice is that methodologies that have traditionally stayed within their disciplinary boundaries, such as Spatial Statistics (Brunsdon & Comber, 2014), have acquired more widespread use. Finally, it is worth pointing out that, because many of the methodologies are computational in nature, there has been an important growth in the use of computational approaches of various kinds. This goes beyond the limited context of using freely available statistical packages such as R or scripting languages like Python, and instead is better represented by researchers applying a variety of computational approaches that support the methodologies mentioned above. Among these one should include concepts and applications of databases, which are of practical relevance when dealing with such large data, as well as more scientifically oriented methods such as agent-based modeling (Epstein & Axtell, 1996).

As a matter of observation, one of the new approaches that has received particular attention in the era of the data revolution has been classification of information (part of Machine Learning). This approach is not new per se but has seen considerable evolution in recent years. Without trying to provide a full or even systematic explanation about why this has happened, part of this tendency has to do with three of the features that come with big data, namely its size; being gathered for purposes not always aligned with its uses in development; and mere novelty.

First, size contributes in various ways to this tendency to classify. On the one hand, data is abundant, hence providing confidence that classification can actually be effective. Additionally, to look for signal when many data features are captured, it is necessary to organize information, leading to classification strategies. Finally, classification as a goal in itself is important, which is illustrated in the various projects that have addressed questions

such as poverty identification in low-income countries (Blumenstock et al., 2015a).

The second point mentioned, that big data is not usually gathered for the purposes for which it ends up being used, contributes to a desire to classify because in order to understand the population sample, one must try to uncover its characteristics. In big data, many kinds of people, institutions, or events get captured, and thus selecting a subset of the data to provide insights requires a kind of pre-processing achieved by classification.

The third point, novelty, reflects the fact that much of the information captured in big data has not been available before, leading to a natural curiosity to learn what is happening in those data and who or what generate them. This scenario, again, encourages classification.

In the future, one would expect the tendency to classify to continue to develop. However, the many other applications that can be given to those data will also continue to evolve. Returning to the main plot of our story, we now discuss in more detail the general methodologies mentioned above.

12.5.1 Common Methodological Frameworks for the Analysis of Big Data

Having described part of the methodological landscape, we can now focus on individual methodologies and *what* they actually do. We provide some very brief summaries of those methods within their scope of application for development. To make these brief descriptions more tangible, we draw on the work done by the United Nations Development Group that has offered an interesting selection of questions on human development and the "Big Data methods" that have been deployed to address them (see (United Nations Development Group, 2017) as well as our Table 12.1 which expands on the referenced document).[2]

12.5.1.1 Traditional Statistical Methods

Classic methods frequently applied in development research today continue to be of fundamental relevance as part of the data revolution. Application of any of the classic techniques such as summary statistics reporting, use of regressions to uncover relations between variables, calculation of correlations, a variety of statistical tests relevant for hypothesis testing, etc. are all encountered frequently in research and applications of Big

[2] We note that the same table is presented in Chapter 11 (reviewing types of big data); there a column referring to the data sources can be found, whereas here that column is replaced by one that makes reference to the methods for the analysis of those data

Table 12.1 Big Data methods used to address the Sustainable Development Goals.

SDG	Theme	Example Research	Methods	References
1	No poverty	Spending patterns as proxies for income	TrS, TS, ML	(Decuyper et al., 2014; Blumenstock et al., 2015b)
2	No hunger	Food-price tracking	TrS, TS SS	(Kim et al., 2017; Zeug et al., 2017)
3	Health	People's movements to predict the spread of infectious diseases	NLP, Net, TS	(Brownstein et al., 2008; Eagle et al., 2009; Salathé et al., 2012)
4	Education	Drivers of student drop-out	ML	(Marquez-Vera et al., 2016; Pérez et al., 2018)
5	Gender equity	Spending and financial activity by gender	TrS, TS, ML, SS	(Martinez et al., 2016; Aron, 2018)
6	Sanitation	Access to clean water	TrS	(Koehler et al., 2015)
7	Clean Energy	Optimization of production and distribution	TrS, Net	(Shupler et al., 2021)
8	Decent Work	Micro-level activity like spending, remittances	TrS, Net	(Ansón & Helble, 2013; Hristova et al., 2016)
9	Infrastructure	Public Transport quality and improvement	TrS, SS	(UN Global Pulse, 2017)
10	Reduce Inequality	Track patterns of discrimination	TrS, ML	(Menon et al., 2017, 2018, Chapter 8 by Hidalgo-Sanchis)
11	Sustainable Communities	Land encroachment on public land	SS, Im	(Bratley & Ghoneim, 2018; Patel & Soni, 2022)
12	Responsible Consumption	Transition to efficient products	TrS, SS, TS	(Zhang & He, 2021)
13	Climate action	Track deforestation	TrS, ML, Im	(Achard et al., 2002; Hansen et al., 2013)
14	Marine life	Track illegal fishing	TrS, ML, SS	(Kurekin et al., 2019)
15	Life on land	Tracking on-land events in real time	TrS, TS, ML NLP, SS	(Kibanov et al., 2017)
16	Justice and Institutions	Track governance, rights, public opinion	ML	(Chapter 8 by Hidalgo-Sanchis)
17	Partnerships for Goals	Connecting data for real-time holistic picture	Im	(K. Anderson et al., 2017)

Note: The table is partly an expansion of the table in the guidance note by the United Nations Development Group (2017). TrS = Traditional Statistical Methods, TS = Time Series Methods, SS = Spatial Statistics, NLP = Natural Language Processing, ML = Machine Learning, Net = Networks, Im = Image Analysis.

Data in Development. Clearly, these techniques are relevant because of the probabilistic context already explained above in section 12.3. In other words, in most cases, the most precise quantitative understanding we can have about the systems captured by Big Data is probabilistic.

Time series analysis can also be considered part of these traditional methods. Techniques in this area are well-known and include, among others, temporal (moving) averages, determination of trends and seasonal (periodic) signals, temporal correlations including time-lagged autocorrelations, forecasting, and many other approaches. These tools remain relevant and applicable to big data, albeit with small adjustments. The connection between probability and time series can originate in several ways. At an intuitive level, if the state of a system makes a change to a new state at the next measurement time point, and this new state cannot be known in advance, one can think of these successive measurements as if they are trials in a chance experiment (e.g., what will the GDP of a country be in two consecutive years?). In this sense, probability seems directly applicable. However, this setting is not identical to simple probabilities because measurements in consecutive time points are generally not unrelated to each other. This means that while many probability concepts apply, they must be used in a different way.

To draw a boundary around what we call traditional statistical methods is slightly hard as it encompasses considerable ground. As a guide, one can think of it as the material found in introductory university classes in statistics or econometrics. There is an abundance of references available for these topics, including some freely available online resources. Among the classic textbooks in introductory statistics, one can find Freedman et al. (2007), and a good freely available alternative is Navarro (2018). For those with econometric preferences, a classic textbook is Wooldridge (2012) and a very recent free source is Hanck et al. (2021). For more dedicated time series analysis, a good freely available source is Hyndman & Athanasopoulos (2018).

12.5.1.2 Machine Learning and Other Methods Connected to Artificial Intelligence

Because of the typical sizes of big data, being able to analyze them semi- or fully-automatically is of critical value, as discussed above in section 12.4. The word learning in Machine Learning is an appropriate aspirational goal for these techniques which try to begin with information and learn patterns from it. Machine learning is built on both statistical and machine computation techniques. Statistics provides the quantitative foundations that can uncover signal from noisy data. Machine computation offers the speed necessary to explore the data automatically and systematically in as many ways as possible to attempt to uncover signal.

The growth of Machine Learning has occurred very rapidly over the last couple of decades, and the techniques have become widely used in both academia and the practitioner world (including the consultancy sector where the term "People Analytics" has emerged in reference to Machine Learning of data connected to human activity). As mentioned before, classification is one of the most prominent areas of application of Machine Learning. This is a problem that can be thought of as mapping the details of the probabilistic context of a situation. What are the features that explain the situation of interest? How does one identify them most effectively? Questions in this spirit are key. The specific methods applied draw on a considerable literature that includes neural networks, dimensional reduction, clustering techniques, feature extraction, and many others. This is also a rapidly evolving field with results emerging at a staggering pace.

Machine Learning can be thought of as being a subset of the field of Artificial Intelligence (AI) which, as its name suggests, attempts to mimic intelligence in terms of a variety of tasks such as gathering information, making optimal decisions, etc. It is not uncommon for both terms to be used interchangeably, but one should keep in mind that ML is narrower in scope. In addition, there is a certain difficulty in always drawing distinctions between ML and AI in that the choice of what is described as AI changes over time, with a tendency to relabel those tasks that at one point were considered challenging but have become routine over time (such as speech recognition). Other large families of methods grouped under the AI umbrella include Neural Networks, Deep Learning, and Natural Language Processing (below). For the purposes of our discussion, what is perhaps most consequential in deciding whether or not to label a problem as being addressed by AI or not is the type of question that any data analysis is meant to address. Concretely, any problem currently considered part of the corpus of Machine Learning, Natural Language Processing, Neural Networks, or any of the related techniques will be less likely to be labeled as being a problem for AI; usually, only problems that are still unresolved tend to be labeled as relevant to AI.

There are well-established references in the ML field which help understand the foundations and some of the most common methods. However, because of its rapidly evolving nature, new references are also emerging. The book by Mitchell (1997) is a great source. A more recent and freely available source is Daumé III, 2017.

12.5.1.3 Natural Language Processing and Large Language Models

The ability to study written information in a fast and automated way opens the door to the extraction of vast amounts of data from text. Since the 1950s,

there has been interest in this goal, running alongside the gradual introduction of the digital computer. Such efforts have evolved in various directions leading to the birth of various areas of research. Natural Language Processing (NLP), one of these products, analyzes text with the purpose of extracting information from written material to make it into quantifiable structured information.

The probabilistic context in which NLP operates usually takes a couple of forms. In one case, the sample space is related to words associated within a document, with probabilities estimated for those word relations. The probabilities then provide information to analyze a document—sometimes to try to understand the document itself, other times to compare it against other documents. Another common version of probability for natural language processing comes from annotated data where words are associated with separate ideas, such as opinions or preferences, possibly gathered directly from people who may be surveyed to offer such impressions on words. This is the basis for so-called sentiment analysis. Note that NLP in any of its forms can also play a role in the process of converting (i.e., pre-processing) unstructured information into structured data, an aspect of data handling that is of critical practical value. Natural Language Processing can also be thought of as a subset or Artificial Intelligence. A good reference for NLP is Bird (2009) and a recent openly accessible source is Eisenstein (2019).

Most recently, an explosion of interest has emerged around so-called Large Language Models (LLMs), based on the neural network architecture called transformer models (Vaswani et al., 2017). These models, embodied in both general and domain-specific versions, are able to interact with written language in a way that closely mimics human discussion. These tools are capable of extracting meaningful information from text, albeit with some level of error and even so-called hallucinations. The importance of these models is that they can drastically simplify NLP tasks and provide outputs specified by a user (structured information, labeling, etc.). Although their impacts are only beginning to be felt, it is extremely likely that the use of text information will undergo revolutionary changes in the near future.

12.5.1.4 Network Science

In many situations, the entities one wants to study are many in number and are also selectively interrelated to each other in non-trivial ways. For example, in social groups where complex social interactions may take place, individuals do not all relate to each other, and those that do relate to one another exercise a form of choice by which some relationships receive a great deal of attention whereas others are serviced at a bare minimum of effort. Such

heterogeneity can be observed in many other contexts such as trade relations between people and/or businesses, suppliers of goods and services, various institutions and organizations, and even countries.

The field of network science (which goes beyond just social network analysis) is concerned with the study of information that interrelates in a networked form (Newman, 2018). An outgrowth and simultaneous convergence of many disciplines has led to the creation of this multidisciplinary field that focuses on the system information that is encoded in the structure of interrelations of that system. For example, in network science a network that displays a whole spectrum of connectivities that range from marginally connected entities to entities that connect a substantial fraction of a network are known to have very different properties than networks with more homogeneously distributed connectivities. Results like these, which hinge not on detailed parts of the network but instead capture the overall structure, are typical of network science. The study of social networks (Wasserman & Faust, 1994; Jackson, 2011; Newman, 2018; Borgatti et al., 2018), networks of infrastructural systems (Buldyrev et al., 2010), distribution or supply chain networks (Surana et al., 2005), institutional networks (Wasserman & Faust, 1994), interactions between Sustainable Development Goals (SDGs) (Klemun et al., 2023), and a number of other examples offer relevant information to issues of Human Development, and are receiving due attention in the literature. A freely available source for a general understanding of network science is Coscia (2021).

12.5.1.5 Spatial Statistics

Spatial location is in many cases a critical component of social and economic phenomena. The advantages that one location has over another (e.g., access to particular resources, better climate, etc.) are manifestations of the importance of space and how it is able to influence outcomes relevant for development. However, just as human-related outcomes are not entirely certain, neither are those associated with space.

While spatial analysis has formed part of development research, with the emergence of Big Data, the study of spatial variation has become more common, and because of the probabilistic nature of spatial outcomes mentioned above, spatial statistical methods have emerged as important tools. Probability in this context usually involves comparisons with baseline models that capture complete spatial randomness, where space plays no special role. Classic references on the subject are abundant; a referential source is Cressie (2015) and a good freely available alternative is De Smith et al. (2007).

12.5.1.6 Databases

The size and complexity of big data requires efficient methods to handle it. In both computer science and engineering, database software and theories have advanced considerably to meet this need. While this is not fundamentally statistical methodology that provides insights, many tasks that involve the analysis of big data are hard to complete without the efficient handling of databases and the associated query languages. The programming model called Map Reduce and its implemented versions (such as Hadoop) exist for this purpose. They basically efficiently distribute both calculations and information over multiple processors and drives, making feasible otherwise impossible computational tasks. A useful openly available resource can be found here (Leskovec et al., 2023).

12.5.1.7 Data Engineering and Data Scraping

A great deal of big data is in fact openly available. This includes information such as communication and transactions on digital platforms (e.g., social media). In the ideal case for the researcher, data is available in straightforward ways so that its acquisition involves simple steps like the download of a file, or a dump of data produced by the data generator or holder. In slightly more complicated situations, it is common to find data through so-called application programming interfaces (APIs) which offer a set of simple instructions that allow for data to be downloaded in some controlled way, containing a selection of desired information that the API can organize from the original data and then transmit.

There are other situations, however, that require more complicated approaches. For example, it is possible to monitor many websites simultaneously in continuous time in order to harvest information, effectively creating big data in the process. This practice is called data scraping and, as described in Marres & Esther Weltevrede (2013), constitutes a socio-technical device because beyond the actual acquisition of data, there are choices about what is captured that the researcher has to make. One important consideration about such scraping is that, in many cases, it contains so-called "live" information, allowing the study of current events. Scraping should be carried out in accordance with ethical standards and respecting the terms and conditions set by the platform.

Considered as a method, data scraping is similar in spirit to other computational techniques. As in the case of databases, it facilitates tasks that would be impossible without the computer. On the other hand, it also imparts a dimension to the data that is different because, contrary to database methods, it is not just concerned with manipulation of data, but it is also responsible for its

creation, in terms of organizing information that can become the source of subsequent analysis.

Alongside data scraping, there is the process of data engineering. This generally refers to the collection of steps needed to take data that may be structured but imperfect (having errors, undesired structure, or other impediments to analysis) and converting them into purified databases that are much more susceptible of efficient study. A great recent reference book is Reis (2022).

12.5.1.8 Modeling

In situations where explanations are being sought, it is very useful to consider modeling techniques such as agent-based models, stochastic models, or systems dynamics, to name a few. In these methods, a mechanism, or a set of them are assumed to explain the phenomena under study.

To illustrate, imagine that local employment rate in a city emerges as a consequence of individual decisions made on the basis of the salaries offered, unemployment insurance, and cost of living. One can then compute a simulated "city" in which only these three factors operate, leading to an overall employment rate. Modeling approaches are related to hypothesis testing in that they can create a chosen probabilistic context against which to compare reality (Epstein & Axtell, 1996).

Some other uses of modeling include policy experiments and precise estimation in epidemiology. A recent framework combines agent-based modeling with open spending data and development indicators to perform impact evaluation in the multidimensional and interdependent context of the SDGs (Guerrero & Castañeda, 2024).

12.5.2 Academic or Research-oriented versus Practitioner Perspectives

As a final point of discussion, it are of great relevance to keep in mind how data are seen, depending on who are using them. This chapter is written with a broad audience in mind and should therefore make distinctions between readers in the academic or research-oriented space and the practitioner space.

It is fair to say that, in either of these spaces, the goal of understanding and engaging with the problems of development partly through the use of data are still a common denominator. However, the specifics become quite different if the goal is to deploy or put something in practice as soon as possible or if,

instead, one is looking for patterns or a causal mechanism that is not intended to be immediately exploited in the field.

Academics and researchers, by the nature of the task, require more time and greater latitude to explore questions than practitioners. Consequently, the application of techniques that are more complicated and/or lengthy will tend to be acceptable by those in research roles, but less relevant to practitioners. Although the practical drivers for this difference are easy to understand, one must keep in mind their advantages and disadvantages.

12.6 Algorithmic Bias and Limitations in Methodology

Although this chapter is dedicated to the description and explanation of the many methods that are used to study large data in the context of development, it would be a glaring oversight not to comment on the problem of algorithmic bias and related technological limitations that make it difficult for the quantitative analysis discussed here to achieve its intended goals.

There are various aspects to this problem that one must consider. First, data themselves are hardly ever unbiased. Sometimes certain groups are omitted from data collection, other times the proper features that characterize them are not gathered. And in many cases, especially when data are not collected for the purposes of well-designed research, the data that become available are merely driven by the market associated with those interested in the tool or setting associated with those data. This data sampling issue is discussed in Chapter 11 on data types.

A second problem arises from the algorithms and methods used to analyze the data. Methods used for classification (such as those mentioned above in regards to ML) will likely be affected by biases at the point of training, leading to biased conclusions in their classification schemes. Poor algorithm design is another important issue that can generate bias even in situations where data have been properly sampled.

Another aspect of the problem is the inherent cycle connected to the generation of some types of big data. Part of such data comes from monetized environments that collect large fractions of information from people using a service for a purpose other than proper data collection. This means that the combination of algorithms and data are purposefully created in order to serve this financial goal. Therefore, reliance on the data generated in such settings is fundamentally challenging when it comes to algorithmic bias.

The increasing awareness of this problem notwithstanding (Dalton & Thatcher, 2014; O'Neil, 2016; Noble, 2018; Benjamin, 2019;

D'ignazio & Klein, 2020; Graham & Dittus, 2022; Sekara et al., 2023), governments are attempting to address such issues with new regulatory frameworks such as the European Union's General Data Protection Regulation (adopted in 2016 and enforced in 2018) and the Artificial Intelligence Act (proposed in 2021). However, much more needs to be done, particularly at the global scale so that a level playing field can emerge across the different spaces around the globe in which the data revolution operates.

12.7 Final Thoughts

The identification of an appropriate method requires that we address two main points:

1) What is the question we want to resolve?
2) What are the data we have?

Besides some obvious use cases, it is not feasible to provide a general road map on how to pick methods. Some off-the-shelf techniques can be appropriate in some cases. Sometimes a technique requires development through minor modifications of existing techniques. Yet other cases require translation of one problem to another because the connections were previously unidentified. In the most challenging situations, new techniques may be needed. In the end, expertise is required in order to critically evaluate and execute these steps. Recognizing the similarity of a specific problem to another one that has already been addressed always helps. Previous literature on development and big data provides a useful guide for the approaches taken in the past. As a practical matter, both academics and practitioners should always be mindful of the literature and the approaches that have been taken in the past. This advice is not just a formality or a matter of giving proper credit or priority, but it is also a practical matter: the ongoing data revolution, when seen in the context of development, is an evolving field where methods and questions change rapidly. Ignoring this information can therefore lead to considerable waste when ongoing or recently published information may offer the most effective strategies by which to tackle a problem.

In summary, in this chapter we have discussed the theoretical and practical aspects of the analysis of big data. Theoretically, we explained the underlying probabilistic setting that describes data and affects how one should analyze them. At the practical level, we described how big data analysis requires statistical tools, the ability to find actionable patterns in the data,

and computational tools, which make it possible for the data to be accessed and used in order to extract analytical insights on the topic of interest. Our main intention has not been to provide an exhaustive review, but rather offer a quick entry point to some of the most frequently used terminology and methods in the field, lowering the entry requirements for those new to the area and offering a simple summary to those already involved in this line of work.

References

Achard, F., Eva, H. D., Stibig, H.-J., Mayaux, P., Gallego, J., Richards, T., & Malingreau, J.-P. (2002). Determination of deforestation rates of the world's humid tropical forests. *Science (New York)*, *297*(5583), 999–1002. https://doi.org/10.1126/science.1070656

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., & Friedl, L. (2017). Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-Spatial Information Science*, *20*(2), 77–96. https://doi.org/10.1080/10095020.2017.1333230

Anderson, R., Perrier, T., Pervaiz, F., Sisouveth, N., Kumar, B., Phongphila, S., Rahman, A., Dhiman, R., & Newland, S. (2014). Supporting immunization programs with improved vaccine cold chain information systems. *IEEE Global Humanitarian Technology Conference (GHTC 2014)*, 215–222. https://doi.org/10.1109/GHTC.2014.6970284

Ansón, J., & Helble, M. (2013). A gravity model of international postal exchanges. In M. A. Crew & P. R. Kleindorfer (Eds.). *Reforming the Postal Sector in the Face of Electronic Competition* (pp. 36–47). Edward Elgar Publishing.

Aron, J. (2018). Mobile money and the economy: A review of the evidence. *The World Bank Research Observer*, *33*(2), 135–188. https://doi.org/10.1093/wbro/lky001

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code* (1st edition). Polity.

Bennett, R., Alberti, G., Çibik, A., Eremenko, T., Formosa, S., Formosa-Pace, J., Jiménez-Buedo, M., Lynch, K., Salazar, L., & Ubeda, P. (2022). Bringing about the data revolution in development: What data skills do aspiring development professionals need? *Journal of International Development*, *34*(7), 1381–1397. https://doi.org/10.1002/jid.3642

Bird, S. (2009). *Natural Language Processing with Python* (1st ed.). O'Reilly.

Blumenstock, J., Cadamuro, G., & On, R. (2015a). Predicting poverty and wealth from mobile phone metadata. *Science (New York)*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Blumenstock, J., Cadamuro, G., & On, R. (2015b). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing Social Networks*. SAGE.

Bratley, K., & Ghoneim, E. (2018). Modeling urban encroachment on the agricultural land of the Eastern Nile Delta using remote sensing and a GIS-based Markov Chain Model. *Land*, *7*(4), Article 114. https://doi.org/10.3390/land7040114

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLOS Medicine*, *5*(7), 1–6. https://doi.org/10.1371/journal.pmed.0050151

Brunsdon, C., & Comber, L. (2014). *An Introduction to R for Spatial Analysis and Mapping*. Sage.

Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., & Havlin, S. (2010). Catastrophic cascade of failures in interdependent networks. *Nature*, *464*(7291), Article 7291. https://doi.org/10.1038/nature08932

Coscia, M. (2021). *The Atlas for the Aspiring Network Scientist*. Michele Coscia. https://www.networkatlas.eu/

Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons.

Dalton, C., & Thatcher, J. (2014). *What Does A Critical Data Studies Look Like, And Why Do We Care?* https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

Daumé III, H. (2017). *A Course in Machine Learning*. http://ciml.info/

De Smith, M. J., Goodchild, M. F., & Longley, P. (2007). *Geospatial analysis: A comprehensive guide to principles, techniques and software tools*. Troubador publishing ltd. https://spatialanalysisonline.com/

Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J.-M., Krings, G., Gutierrez, T., Blondel, V. D., & Luengo-Oroz, M. A. (2014). *Estimating Food Consumption and Poverty Indices with Mobile Phone Data*. arXiv. https://doi.org/10.48550/ARXIV.1412.2595

D'ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Eagle, N., Pentland, A. (Sandy), & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278. https://doi.org/10.1073/pnas.0900282106

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089. https://doi.org/10.1126/science.1243089

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT press.

Epstein, J. M., & Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. W. W. Norton & Company.

Graham, M., & Dittus, M. (2022). *Geographies of Digital Exclusion: Data and Inequality*. Pluto Press. https://cir.nii.ac.jp/crid/1130293004051685790

Guerrero, O., & Castañeda, G. (2024). *Complexity Economics and Sustainable Development: A Computational Framework for Policy Priority Inference*. Cambridge University Press.

Hanck, C., Arnol, M., Gerber, A., & Schmelzer, M. (2021). *Introduction to Econometrics with R*. https://www.econometrics-with-r.org/index.html

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & J. R. G. Townshend. (2013). High-resolution global maps of 21st-century forest cover change. *Science (New York)*, *342*(6160), 850–853. https://doi.org/10.1126/science.1244693

Harris, J. (2022). Mixed methods research in developing country contexts: Lessons from field research in six countries across Africa and the Caribbean. *Journal of Mixed Methods Research*, *16*(2), 165–182. https://doi.org/10.1177/15586898211032825

Hristova, D., Rutherford, A., Anson, J., Luengo-Oroz, M., & Mascolo, C. (2016). The international postal network and other global flows as proxies for national wellbeing. *PLOS ONE*, *11*(6), 1–19. https://doi.org/10.1371/journal.pone.0155976

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts. https://otexts.com/fpp2/

Jackson, M. O. (2011). Chapter 12—An Overview of Social Networks and Economic Applications. In J. Benhabib, A. Bisin, & M.O. Jackson (Eds), *Handbook of Social Economics* (Vol. 1, pp. 511–585). North-Holland. https://doi.org/10.1016/B978-0-444-53187-2.00012-7

Kibanov, M., Stumme, G., Amin, I., & Lee, J. G. (2017). Mining social media to inform peatland fire and haze disaster management. *Social Network Analysis and Mining*, *7*(1), 30. https://doi.org/10.1007/s13278-017-0446-1

Kim, J., Cha, M., & Lee, J. G. (2017). Nowcasting commodity prices using social media. *PeerJ Computer Science*, *3*, e126.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Klemun, M. M., Ojanperä, S., & Schweikert, A. (2023). Toward evaluating the effect of technology choices on linkages between sustainable development goals. *iScience*, *26*(2), 105727. https://doi.org/10.1016/j.isci.2022.105727

Koehler, J., Thomson, P., & Hope, R. (2015). Pump-priming payments for sustainable water services in rural africa. *World Development*, *74*, 397–411. https://doi.org/10.1016/j.worlddev.2015.05.020

Kurekin, A., Loveday, B., Clements, O., Quartly, G., Miller, P., Wiafe, G., & Adu Agyekum, K. (2019). Operational monitoring of illegal fishing in ghana through exploitation of satellite earth observation and AIS data. *Remote Sensing*, *11*(3), 293. https://doi.org/10.3390/rs11030293

Leskovec, J., Rajaraman, A., & Ullman, J. (2023). *Mining of Massive Datasets*. Infolab. http://www.mmds.org/

Marquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Marres, N., & Weltevrede, E. (2013). Scraping the Social? *Journal of Cultural Economy*, *6*(3), 313–335. https://doi.org/10.1080/17530350.2013.772070

Martinez, E. A., Rubio, M. H., Martinez, R. M., Arias, J. M., Patane, D., Zerbe, A., Kirkpatrick, R., Luengo-Oroz, M., & Zerbe, A. (2016). *Measuring Economic Resilience to natural disasters with big economic transaction data*. arXiv. https://doi.org/10.48550/ARXIV.1609.09340

Menon, R., Biswas, A., Armin, S., Quinn, J., & Niesler, T. (2018). Automatic Speech Recognition for Humanitarian Applications in Somali. *Proceedings of SLTU*. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), Gurugram, India.

Menon, R., Saeb, A., Cameron, H., Kibira, W., Quinn, J., & Niesler, T. (2017). Radio-browsing for developmental monitoring in Uganda. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5795–5799. https://doi.org/10.1109/ICASSP.2017.7953267

Mitchell, T. M. (1997). *Machine Learning* (Vol. 1). McGraw-Hill.

Navarro, D. (2018). *Learning Statistics with R: A tutorial for psychology students and other beginners*. https://learningstatisticswithr.com/

Newman, M. (2018). *Networks*. Oxford University Press. https://doi.org/10.1093/oso/9780198805090.001.0001

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://nyupress.org/9781479837243/algorithms-of-oppression

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Patel, A., & Soni, J. B. (2022). Bhoomi Prahari—e-governance tool for monitoring encroachment on government land using mobile and GIS technology. *International Journal of Information Technology*, *14*(3), 1565–1573. https://doi.org/10.1007/s41870-021-00674-2

Pérez, B., Castellanos, C., & Correal, D. (2018). Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study. In A.D. Orjuela-Cañón, J.C. Figueroa-García, & J.D. Arias-Londoño (Eds), *Applications of Computational*

Intelligence (pp. 111–125). Springer International Publishing. https://doi.org/10.1007/978-3-030-03023-0_10

Poggio, T., & Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the AMS*, *50*(5), 537–544.

Reis, J. (2022). *Fundamentals of Data Engineering: Plan and Build Robust Data Systems*. O'Reilly.

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., & Vespignani, A. (2012). Digital epidemiology. *PLOS Computational Biology*, *8*(7), 1–3. https://doi.org/10.1371/journal.pcbi.1002616

Sekara, V., Karsai, M., Moro, E., Kim, D., Delamonica, E., Cebrian, M., Luengo-Oroz, M., Jiménez, R. M., & Garcia-Herranz, M. (2023). *Are Machine Learning Technologies Ready to Be Used for Humanitarian Work and Development?* (arXiv:2307.01891). arXiv. https://doi.org/10.48550/arXiv.2307.01891

Shupler, M., O'Keefe, M., Puzzolo, E., Nix, E., Anderson de Cuevas, R., Mwitari, J., Gohole, A., Sang, E., Čukić, I., Menya, D., & Pope, D. (2021). Pay-as-you-go liquefied petroleum gas supports sustainable clean cooking in Kenyan informal urban settlement during COVID-19 lockdown. *Applied Energy*, *292*, 116769–116769. https://doi.org/10.1016/j.apenergy.2021.116769

Smyth, P. (2000). Data mining: Data analysis on a grand scale? *Statistical Methods in Medical Research*, *9*(4), 309–327. https://doi.org/10.1177/096228020000900402

Surana, A., Kumara, S., Greaves, M., & Raghavan, U. N. (2005). Supply-chain networks: A complex adaptive systems perspective. *International Journal of Production Research*, *43*(20), 4235–4265. https://doi.org/10.1080/00207540500142274

UN Global Pulse. (2017). *Using Big Data Analytics for Improved Public Transport*. UN Global Pulse. https://pulselabjakarta.org/assets/uploadworks/2018-08-02-13-21-05.pdf

United Nations Development Group. (2017). *Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda*. United Nations Development Group. undg.org/document/data-privacy-ethics-and-protection-guidance-note-on-big-data-for-achievement-of-the-2030-agenda/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815478

Wooldridge, J. (2012). *Introductory Econometrics: A Modern Approach* (5th ed.). South-Western College Pub.

Zeug, H., Zeug, G., Bielski, C., Solano-Hermosilla, G., & M'Barek, R. (2017). *Innovative food price collection in developing countries: Focus on crowdsourcing in Africa. 28247.* http://publications.europa.eu/publication/manifestation_identifier/PUB_LBNA28247ENE

Zhang, G., & He, B.-J. (2021). Research on the Global Green Market Based on Big Data. In D. La Rosa & R. Privitera (Eds), *Innovation in Urban and Regional Planning* (pp. 139–148). Springer International Publishing.

PART IV

ISSUES OF ETHICS, JUSTICE, AND LEGAL FRAMEWORKS RELATED TO BIG DATA ANALYSIS IN HUMAN DEVELOPMENT

13

Data Justice for International Development

A Structural Model and Manifesto

Richard Heeks

13.1 Introduction

We are moving towards an era of "data-intensive development," involving "the growing presence and application of data in the processes of international development" (Heeks, 2017a). For example, there is growth in availability and use of big data (Hilbert, 2016), and there is growth in availability and use of open data (WWW, 2017). With this expansion of data-intensive development have come a number of development benefits: faster and better decisions that are facilitating improved development outcomes in health, agriculture, urban planning, etc. (Cartesian, 2014; Kshetri, 2014). But there have also been a number of growing concerns about emerging negative impacts: loss of privacy, discrimination, growth in inequality, etc. (Spratt & Baker, 2015; Taylor & Broeders, 2015).

Those concerns have been framed in a number of ways but one strand of work has linked itself to "data justice," that is, "the specification and pursuit of ethical standards for data-related resources, processes and structures" (Heeks, 2017b). This is still a formative topic area with a small number of papers emerging in recent years (e.g., Newman, 2015; Dencik et al., 2016), including one paper specifically addressing data justice in the context of international development (Heeks & Renken, 2018).

This latter paper had a specific argument. Having provided evidence of data injustices occurring in developing countries, it identified three mainstream perspectives on the converse, "data justice" (Heeks & Renken, 2018:92):

Richard Heeks, *Data Justice for International Development*. In: *Human Development and the Data Revolution*. Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Richard Heeks (2025).
DOI: 10.1093/oso/9780198879145.003.0014

- "Instrumental data justice means fair use of data; it therefore focuses on the outcome of use of data."
- "Procedural data justice means fair handling of data" along all parts of the information value chain (see Figure 13.1; Heeks, 2018).
- Distributive rights-based justice encompasses rights of privacy, access, ownership and representation; the enactment of which shapes distribution of data resources.

However, the Heeks & Renken (2018) paper also argued that these three mainstream views of data justice had a number of shortcomings. In particular, a failure to "encompass the social structures which at least partly determine data uses, processes, distributions, and rights" (Heeks & Renken, 2018: 96). It further argued that these structures needed to be encompassed and that "the foundation of data justice must be structural data justice, which we can define as 'the degree to which society contains and supports the data-related institutions, relations and knowledge systems necessary for realization of the values comprised in a good life.'"

Heeks & Renken (2018: 97) propose three different ways in which this idea of structural data justice could be conceptualized:

1. A connectivist view, building on "the work of Iris Marion Young who takes a network view of social structure—what she calls her 'social connection model'—seeing individuals occupying particular social positions, and that structure 'consists in the connections among these positions and their relationships, and the way the attributes of positions internally constitute one another through those relationships' (Young, 2006: 112)."
2. A critical view, looking at the competing interests that intersect around data systems and which have different relative degrees of power with which to shape the design and outcome of those systems. This would first understand interests, "particularly the conflicting interests relating to data rights, processes, distribution and uses," and would understand structure as "the mechanisms of power through which those interests come to either strengthen or weaken."
3. A capabilities view building from the work of Amartya Sen. This understands justice to mean "increasing the freedoms of individuals to be and to do what they value in life" (ibid.: 98); those freedoms being termed "capabilities." It would see structure as the social context of institutions, hierarchies, power relations, etc. that enable and constrain the realization of individual data-related freedoms.

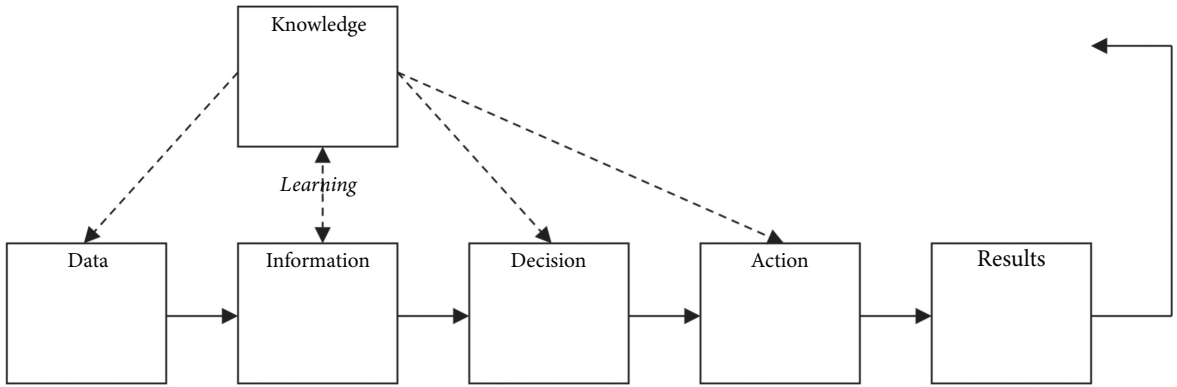


Figure 13.1 The information value chain.

Reproduced with permission from Heeks, R. (2018) *Information and Communication Technology for Development (ICT4D)*, Routledge, Abingdon, UK.

Note: The core of the information value chain (solid lines) is the processes by which data is processed into usable information which is then used to feed decision making and action, which leads to development results. Those results in turn feedback to create new data and new knowledge. The dashed lines indicate the supervisory role of knowledge: guiding and shaping the capture of data, its processing into information, decision making, and action. Information may sometimes not be immediately used in decisions but, instead and via learning, create new knowledge.

Readers are referred to Heeks & Renken (2018) for further detail. However, that paper left a knowledge gap because it was developed solely on the basis of secondary data, and its models for structural data justice were propositions rather than empirically derived frameworks. This current chapter reports on a set of field studies that can help to address this knowledge gap, with the purpose of the chapter being to take one of the models of structural data justice proposed by Heeks & Renken (2018) and to develop it in light of evidence from those field studies. The main contribution of the chapter is thus the production of a new, empirically based model of structural data justice.

The chapter proceeds as follows. First, a short evidence section will outline the field studies undertaken. Then the main body of the chapter reviews a basic model of structural data justice and steadily refines it on the basis of field study evidence, which is presented as a set of three tests for a framework of structural data justice. The chapter then reflects on further elements in Heeks & Renken (2018): the extent to which the revised model captures other proffered conceptualizations, and finally the implications of the revised model for practice; specifically in deriving a proposed "Data-Justice-for-Development Manifesto."

13.2 Field Study Evidence

The evidence reported below derives from four field studies of data-intensive development, commissioned and co-developed by the author. These are:

1. A study of big data in an Indian state electricity corporation. The study was led by Ritam Sengupta, was coordinated by Sumandro Chattapadhyay of the Centre for Internet and Society, Bengaluru, India, and is published as Sengupta et al. (2017) and in Heeks et al. (2021). Losses in electricity distribution—both technical losses plus non-payment of bills—led the corporation to install digital meters (some online, some offline that had to be read via human intervention) for all consumers and electricity transformers. The terabytes of data generated are used to some extent in billing, rectification of faults, and planning improvements to the electricity network.
2. A study of big data in the Bengaluru Metropolitan Transport Corporation (BMTC), which runs the city's 6,000-plus buses. The study was led by Vanya Rakesh, was coordinated by Sumandro Chattapadhyay of the Centre for Internet and Society, Bengaluru, India, and is published as Rakesh et al. (2018) and in Heeks et al. (2021). The BMTC has financed

development of a new big data system, called the Intelligent Transport System (ITS). The main elements of ITS are online, real-time tracking of all buses, and electronic ticketing machines from which all ticketing data is uploaded. The gigabytes of data generated are used to feed a passenger information system that is intended to show where buses are and when they will arrive at bus stops, and to support operational management such as cash management of ticketing. It is the future intention within BMTC that data should be used for tactical management such as rationalization of bus routes and schedules.

3. Two studies that were coordinated and led by Satyarupa Shekhar of the Citizen Consumer and Civic Action Group, Chennai, India, and published as Heeks & Shekhar (2019) which reports, among other things, the experiences of two initiatives that developed data relating to marginalized urban communities. Map Kibera facilitated the digital mapping of Kibera slum in Nairobi by community members, gathering data via on-the-ground surveying of community locations (roads, buildings), assets (schools, clinics) and issues (crime, poverty). Transparent Chennai was a basket of projects that gathered data on various issues in the city: homelessness, availability of public services, vacant land available for development, and services in the tenement blocks provided for evicted slum dwellers. In both cases, data was then visualized via digital and paper maps intended for use in urban planning, particularly by local government.

These four projects reflect a mix of data from formal and informal development settings, and a mix of different data forms and volumes. While they have context specificities, they are representative of some of the key types of data-intensive development project becoming increasingly prevalent in developing countries. They therefore provide a useful basis for exploration of structural data justice in relation to international development.

13.3 Developing a Structural Model of Data Justice

As explained above, Heeks & Renken (2018) proposed a number of possible ways forward for conceptualizing structural data justice. The foundation model used here will be the basic one proposed from critical data studies, reproduced in Figure 13.2.

In this section, we review this model by considering evidence from the field studies. Derived from the studies are three tests for any model of structural

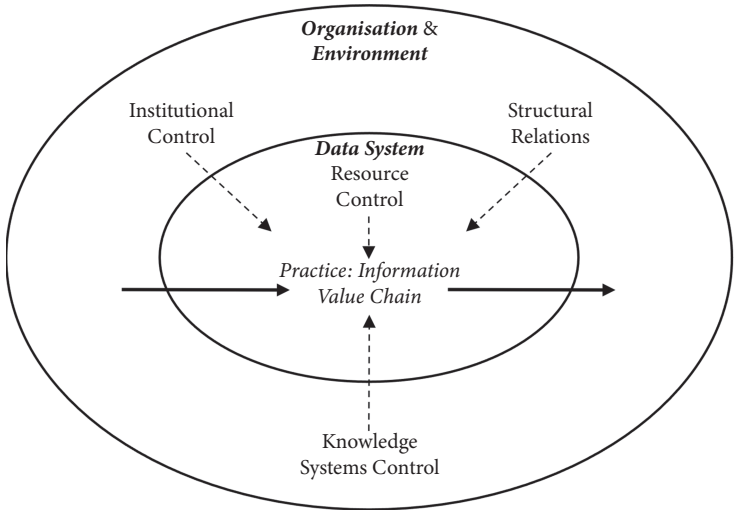


Figure 13.2 Foundation model of structural data justice.

data justice. These are used to develop a more comprehensive—also of course more complex—model of structural data justice.

13.3.1 Test 1: Can the Model Encompass the Impact of Pre-existing Structure?

It is, of course, a core argument of structural data justice that structural features shape other aspects of data justice. Many different definitions of social structure are possible, but the approach taken here sees social structure in terms of power—particularly in terms of "power over," that is, the ability of powerful actors (because of the disposition of social structure) to shape other aspects of data justice. Drawing from the literature on social structure and power (e.g., Hardy & Phillips, 1998; Hearn, 2012), power is seen to come from control over resources and practices, but also from three other things that themselves determine control over resources and practices: control over institutions, position within structural relations, and epistemic control over knowledge systems. It is this view which led to the definition of structure above as "data-related institutions, relations and knowledge systems," given that resources and practice were already encompassed in distributive and procedural notions of data justice.

We can see examples in practice of this structural shaping of data justice. In the case of the Indian electricity corporation, installation of digital electricity

metering for consumers could have led to a sharp reduction in financial losses for the electricity corporation. This would have occurred as accurate data on electricity usage led to accurate billing for all consumers, and thence to more effective collection of amounts owed. This *did* happen in urban areas but not elsewhere. Instead, there was estimated to have been a growth in those not paying or underpaying their electricity bills in rural areas, with potential for inequality given "it was the richer and better-politically-connected sections in [the State's] rural areas—rural industrialists, larger-scale farmers, larger-scale irrigation owners—who were most likely to be appropriating free or low-cost electricity" (Sengupta et al., 2017: 20). The explanation is that this big data initiative was inserted into the heavily politicized nature of electrification in India. Those with political power—politicians with power by virtue of their institutional control and structural position—saw that big data was increasing billing compliance rates in urban areas and was also leading to fewer blackouts and brownouts. This enabled the politicians to drive a large-scale program of electrification in rural areas, with an implicit notion that higher urban revenues could cross-subsidize not just rural expansion but also any rural non-payment. This was a program that they knew would win votes, especially when more powerful consumers were allowed to circumvent the accurate billing that new metered data streams could provide.

In the electricity case, wider structures impacted the way big data was being used (thus linking to procedural data justice) and the outcomes of that data use (thus linking to instrumental data justice). The field studies also exposed instances of wider structures shaping distributive rights-based justice: privacy, access, ownership, and representation. For example, the two sectoral case studies undertaken in India—electricity and transport—involved public sector organizations. They gathered data from citizens (i.e., electricity consumers and bus users) and they gathered data about public services. Yet in neither of those cases to date has the data been made available in some form of open data. This has arisen because such data access goes against the institutional norms of the organizations involved, and more directly because it goes against the interests of the powerful figures within those organizations, who fear the impact of data-driven transparency.

From these examples and others, we could see ways in which the four elements of structure (structural position, and resource, institutional and epistemic control) influence the practices of the information value chain; thus validated the model above. However, we can also see some shortcomings of the model. The influence of structure on procedural data justice is well-reflected. But the influence on instrumental data justice—the outcomes of use of data such as those in the Indian electricity case—is not. And nor is

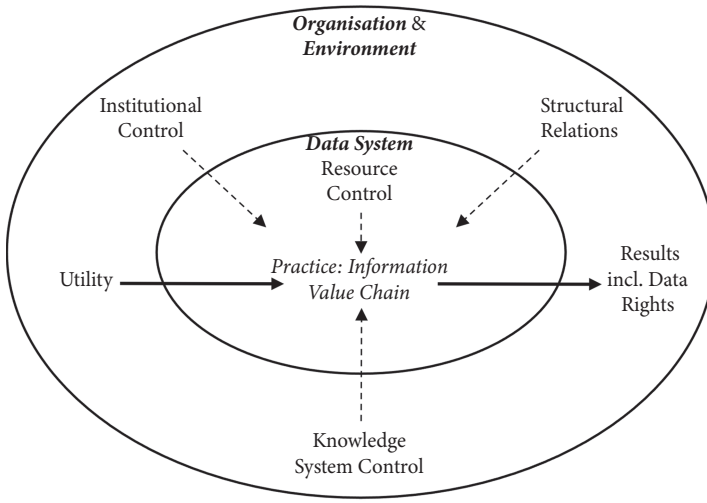


Figure 13.3 First iterative development of structural data justice model.

the influence on distributive rights-based justice such as the right to access. Development results therefore need to be acknowledged as an output from the information value chain. And data rights need to explicitly appear.

Second, structure has to be enacted. In the cases noted above, it was actors with structural power who shaped the data system and its practices. From the structural elements included in the model we can understand *how* they are able to shape data rights, practices, and outcomes through their structural power. But the cases were driven by the particular interests of these actors: the desire to win votes, or the desire to avoid scrutiny which explains *why* they seek to shape data rights, practices, and outcomes. This driving force shapes the information value chains that run and those that do not, and it could be given various names such as interests, incentives, or values. Here, I suggest using the term "utility" to represent the combination of extrinsic incentives and intrinsic motivations and beliefs that shapes the goals that actors value and seek to achieve through the information value chain. These developments are incorporated into the model, as shown in Figure 13.3.

13.3.2 Test 2: Can the Model Encompass the Impact of Data Systems on Structure?

The picture painted so far has been rather static and unidirectional: focusing on ways in which the justice or injustice of wider structures impacts other

aspects of data justice. But structure is not static. It is influenced by data systems and the practices of the information value chain.

For example, the advent of big data in the Indian electricity and transport cases has been associated with a change in structural relations within the respective organizations, and an exacerbation of pre-existing structural inequalities. In both cases, there have been two upward shifts of power towards central management: from labor to management, and from middle to central management. In the electricity corporation, for instance, the introduction of online metering has already led to almost half of meter-readers losing their jobs, and those that remain capture a declining proportion of the corporation's overall data. In the bus transport corporation, "Big data systems enable central management to directly access performance data from the front-line of operations, and to automatically undertake and communicate performance management; obviating the need for intermediating management layers" (Rakesh et al., 2018: 18). As a result, the corporation's divisional offices—the layer that sits between central management and the bus depots—are being closed: representing a significant organizational restructuring.

To take another example, we have seen the way in which all the data initiatives in India have changed the surrounding epistemics. Data systems do this because they create a separate virtual model of the phenomena about which they gather data: the so-called "data double" or, for place-based representations, the "shadow map" (Taylor & Broeders, 2015). This in turn can alter the "imaginaries" of those involved—the mental model and worldviews they have about the phenomena—and the wider discourse about those phenomena. For example, big data in the bus transport corporation has changed the mental model of managers: "the daily operations of the bus fleet and bus crews were largely opaque to management prior to ITS, but they are becoming increasingly visible and thus changing the perceived picture of BMTC that managers hold in their heads" (Rakesh et al., 2018: 17).

In the urban data initiatives, we also found examples of data-enabled epistemic reframing. In the case of Map Kibera, the data gathered from the informal settlement enabled some reorientation of perceptions and discourse among urban planners and officials. They moved somewhat from an epistemic of the troublesome slum-dweller that needed eviction, to a discourse about informal residents who have rights, including a right of access to public services.

Of course, by what they make visible and spotlight, data systems may throw into further shadow that which is *not* visible through digital data. In the bus transportation case, datafication of the buses and ticketing has reinforced

a discourse that focuses on the (negative) actions of the bus crews: speeding, missing bus stops, diverging from the set route by bus drivers, failure to issue tickets, and theft by the bus conductors. Framing these as the key problems of bus transportation takes the discourse away from other issues such as complaints about corruption within operational and strategic management of the corporation. Similarly, in one urban data case, data was gathered on the quality of tenements to which evicted slum dwellers had been relocated. By strengthening the discourse around improvements needed in the tenements, this initiative rendered conditions in informal slums and the process of eviction less visible, and implicitly legitimized eviction as an urban development strategy in Chennai.

The impact of data systems on wider structures is acknowledged in the model through the arrow into the outer domain. However, it would seem better to provide a more explicit recognition of this by changing the links between structure and data practice to two-way interactions. In focusing on the two-way relation between data systems and social structure, the position of data rights comes under further scrutiny. It is not merely that rights are impacted by data systems; they also shape those systems. Indeed, data rights are part institution, part episteme: for instance, the norms and discourse around rights of access to public sector data are part of what is shaping the pressure on various Indian organizations to open up their datasets (albeit, they have resisted so far for the reasons noted above). Finally, given the focus on the language of "epistemic," it would seem simpler to re-phrase "knowledge systems control" to "epistemic control."

We can therefore move to a second iterative development of the model, as shown in Figure 13.4.

13.3.3 Test 3: Can the Model Encompass Datafication?

In the examples given above, the technology and its capabilities play a central role. It is the digitizing capabilities of the online electricity meters that lead to the declining requirement for human meter-readers. It is the datafying capabilities of the Intelligent Transport System that create the data double of the bus system that enables a new imaginary in the minds of managers. Yet these capabilities are not well reflected in the model so far.

To understand these, we can differentiate capabilities into the functionalities and the affordances of data systems. The core functionalities of new data systems can be described in terms of the much-used "3V" data qualities: that

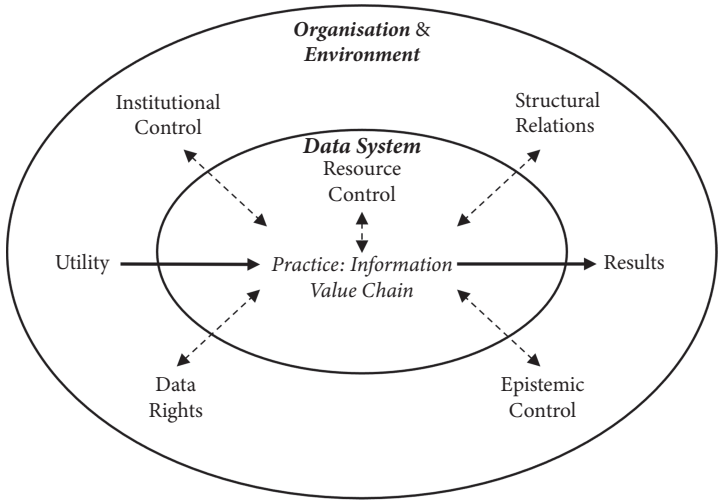


Figure 13.4 Second iterative development of structural data justice model.

they provide a greater volume, velocity, and variety of data than previous systems. "The qualities are inherent functionalities of data. From these qualities, combined with purposive use by individuals or organizations, the following affordances emerge:

- ***Datafication***: an expansion of the phenomena about which data are held. A greater breadth: holding data about more things. A greater depth: holding more data about things. And a greater granularity: holding more detailed data about things. This is accelerated by the second affordance . . . ***Digitization***: not just the conversion of analogue to digital data but the same conversion for all parts of the information value chain. Data processing and visualization for development becomes digital; through growth of algorithms, development decision-making becomes digital; through growth of automation and smart technology, development action becomes digital. Digitization means dematerialization of data (its separation from physical media) and liquification of data (its consequent fluidity of movement across media and networks), which underlie the third affordance . . . ***Generativity***: the use of data in ways not planned at the origination of the data. In particular, data's reprogrammability (i.e., using data gathered for one purpose for a different purpose); and data's recombinability (i.e., mashing up different sets of data to get additional, unplanned value from their intersection)" (Heeks, 2017a, developed from Lycett (2013) and Nambisan (2016)).

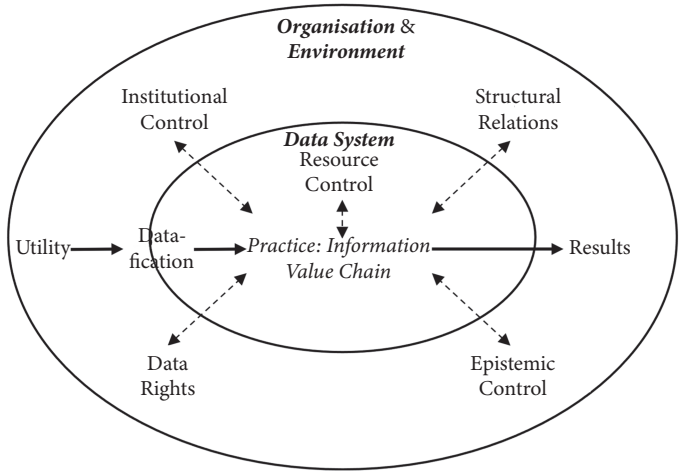


Figure 13.5 Revised structural data justice model.

Without consideration of these affordances, the model of structural data justice would give too little weight to the technology that lies at the heart of both structure and process; would fail to acknowledge the "data-ness" of structural data justice. To understand how to incorporate this, we can use the definition of affordances as "the potential actions an individual or organization with a purpose can undertake with the [*data*] system within the context of the environment within which they function" (Heeks, 2018: 13). With purpose already included via utility, affordances therefore act as a filter between utility and practice; delimiting but also enabling the practices that can be undertaken. For the sake of simplicity, in Figure 13.5, only the first affordance—datafication—is explicitly represented but this can be seen as standing for all the affordances of data systems.

13.4 Discussion

The revised structural data justice model therefore represents the recommended starting point for future research on data and development. It can be used pre hoc to design research, and as a post hoc analytical lens to re-analyze existing cases of data-intensive development. Nonetheless, we can still interrogate the model further.

Having developed the model of structural data justice on the basis of field experience, we can check where the four forms of data justice fit into the model. In doing this—see Figure 13.6—it can be seen that all four

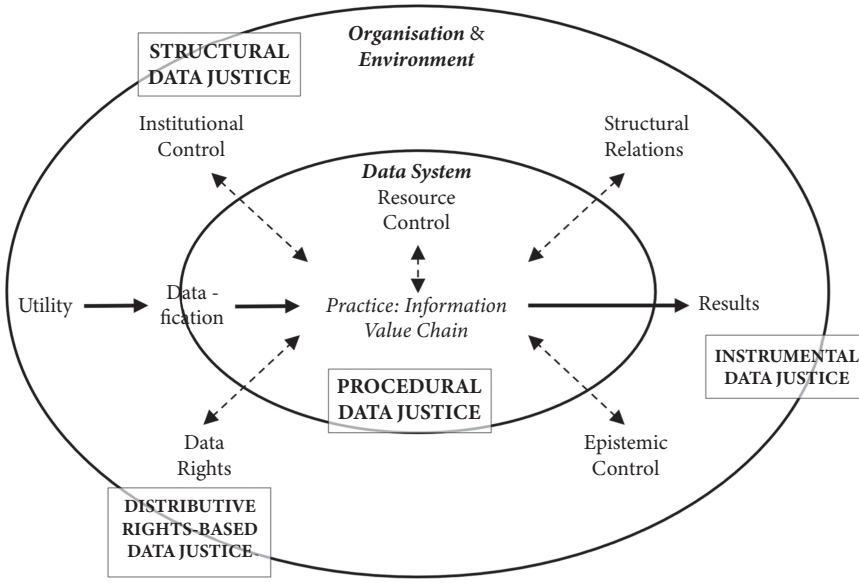


Figure 13.6 Forms of data justice within the revised structural data justice model.

forms are represented, but also that the separation of types of justice may be somewhat artificial. For example, data rights are part of broader social structure, and impacts of data systems at least partly affect broader social structure.

What about the other conceptual routes proposed in Heeks & Renken (2018); how does this model relate to them? In terms of Young's (2006) social connection model, the revised structural data justice model (Figure 13.5) clearly encompasses social connection in its consideration of structural relations. However, it considers this network/positional perspective as only one aspect of social structure, so it is hard to argue that this is a direct operationalization of Young's model.

Such an argument might be easier to sustain in the case of critical data studies. Indeed, the origins of the foundational (Figure 13.2) model were as an illustration of critical data studies. A useful concept here is the idea of "data assemblages," that is, "all of the technological, political, social and economic apparatuses that frame [the] nature, operation and work" of data systems (Kitchin & Lauriault, 2014: 6). One could readily make the case that institutions, structural relations, knowledge systems and rights represent those assemblages. The structural data justice model does not separately identify technology as part of the assemblage, but technological infrastructure can be recognized as both an institutional and structural element, which

provides the datafication and related affordances. Development of the foundational model to include utility does also acknowledge the notion of actor interests and the way in which those are either strengthened or weakened through the operation of data systems; an important focus for critical data studies.

Where the model is less clear is in the idea of conflicting interests: there is no obvious dialectic and yet conflicting interests do emerge from the field studies in relation to the views and interests of differing groups. In implementation of the model, one would therefore need to ensure a critical methodology such as critical systems heuristics which asks: who gets what; who controls what; who does what; who gets affected by the process, and with what justification (adapted from Reynolds, 2014).

13.4.1 Data-Related Agency and Capabilities

As noted in Heeks & Renken (2018), a danger of structural approaches is structural determinism that misses the role of human agency. The incorporation of practice and the presence of decisions and actions within the information value chain mean the revised structural data justice model can analyze human action. But the essence of agency is a freedom and autonomy to act. As an input to the model, this "power to" does indeed derive from control over institutions, resources, episteme, etc., as shown in the revised structural data justice model (Figure 13.5) but it is not specifically identified. And in relation to the impact on agency this is just subsumed within "results": both the agency of those who are directly using the informational or decisional outputs of the data systems (such as electricity corporation managers planning extensions to the distribution network), and those indirectly affected by the emergent decisions and actions (such as tenement residents whose service access is improved).

To bring this into the model, Sen's ideas of capabilities and functionings can be included, as shown in Figure 13.7.

The model remains significantly structural—appropriately so, given the focus here on structural data justice—and more suitable to analyze direct users of data systems than those indirectly affected. Because of that, it might well be more suitable—for those with particular interests in "justice-in-practice" and in justice understood as the expansion of personal freedoms—to take a more capabilities-based approach to conceptualization, such as those proposed by Heeks & Renken (2018: Figure 3) or Taylor (2017: Figure 2).

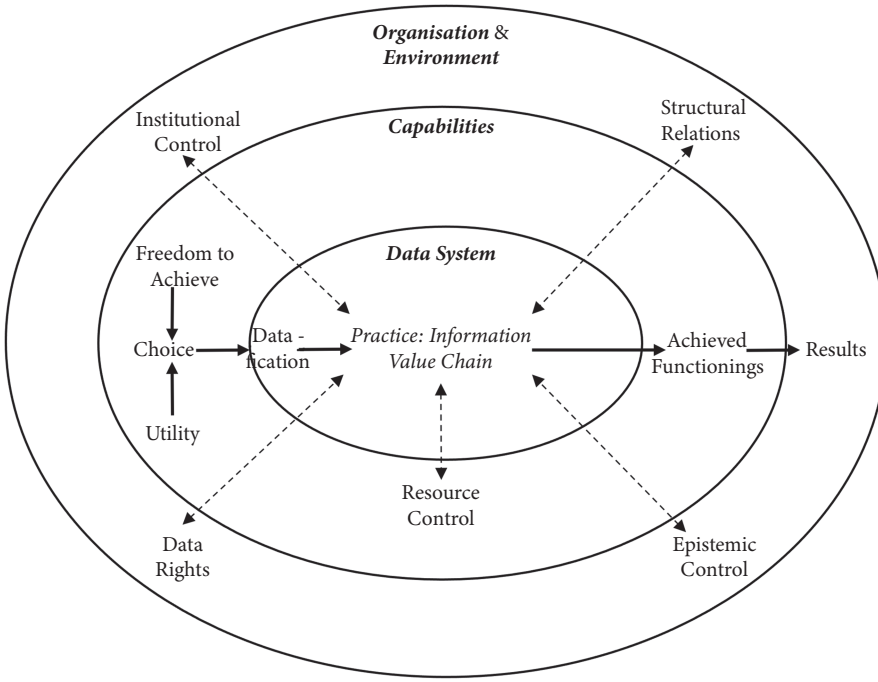


Figure 13.7 Incorporating explicit agency and capabilities into the structural data justice model.

13.4.2 A Data-Justice-for-Development Manifesto

Finally, we can draw from all elements of the structural data justice model—including the other aspects of data justice that it encompasses—to propose a "Data-Justice-for-Development Manifesto":

1. Demand just and legal uses of development data.
2. Demand data consent of citizens that is truly informed.
3. Build upstream and downstream data-related capabilities among those who lack them in developing countries.
4. Promote rights of data access, data privacy, data ownership and data representation.
5. Promote data system outcomes that address international development goals and priorities; including the goals and priorities of data subjects.
6. Support "small data" uses by individuals and communities in developing countries.
7. Advocate sustainable use of data and data systems.

8. Create a social movement for the "data subalterns" of the Global South.
9. Stimulate an alternative discourse around data-intensive development that places issues of justice at its heart.
10. Develop new organizational forms such as data-intensive development cooperatives.
11. Lobby for new data justice-based laws and policies in developing countries (including action on data monopolies).
12. Open up, challenge, and provide alternatives to the data-related technical structures (code, algorithms, standards, etc.) that increasingly control international development.

This can be used as a guiding set of principles for practice—that is, broad policy, organizational strategy, and specific operational planning—aiming to deliver more just development outcomes from data-intensive development.

References

Cartesian (2014). *Using Mobile Data for Development*. Boston, MA: Cartesian.

Dencik, L., Hintz, A., & Cable, J. (2016). Towards data justice? The ambiguity of anti-surveillance resistance in political activism, *Big Data & Society*, 3(2), 1–12.

Hardy, C., & Phillips, N. (1998). Strategies of engagement. *Organization Science*, 9(2), 217–230.

Hearn, J. (2012). *Theorizing Power*. Basingstoke, UK: Palgrave Macmillan.

Heeks, R. (2017a). The affordances and impacts of data-intensive development, *ICT4DBlog*, June 5, https://ict4dblog.wordpress.com/2017/06/05/the-affordances-and-impacts-of-data-intensive-development/

Heeks, R. (2017b). Data justice: Addressing equality, sustainability and transformation through data, presentation at IFIP WG9.4 14th international conference *ICTs for Promoting Social Harmony*, Yogyakarta, Indonesia, May 22–24.

Heeks, R. (2018). *Information and Communication Technology for Development (ICT4D)*. Abingdon, UK: Routledge.

Heeks, R., Rakesh, V., Sengupta, R., Chattapadhyay, S., & Foster, C. (2021). Datafication, value and power in developing countries: Big data in two Indian public service organisations. *Development Policy Review*, 39(1), 82–102.

Heeks, R., & Renken, J. (2018). Data justice for development: What would it mean? *Information Development*, 34(1), 90–102.

Heeks, R., & Shekhar, S. (2019). Datafication, development and marginalised urban communities: An applied data justice framework. *Information, Communication & Society*, 22(7), 992–1011.

Hilbert, M. (2016). Big data for development. *Development Policy Review*, 34(1), 135–174.

Kitchin, R., & Lauriault T. P. (2014). *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work*. Programmable City Working Paper 2, Maynooth, Ireland: National University of Ireland Maynooth.

Kshetri, N. (2014). The emerging role of big data in key development issues, *Big Data & Society*, 1(2), 1–20.

Lycett, M. (2013). "Datafication": Making sense of (big) data in a complex world. *European Journal of Information Systems*, 22(4), 381–386.

Nambisan, S. (2016). Digital entrepreneurship: Toward a digital technology perspective of entrepreneurship, *Entrepreneurship Theory and Practice*, 41(6), 1029–1055.

Newman, N. (2015). *Data Justice: Taking On Big Data As An Economic Justice Issue*. New York: Data Justice.

Rakesh, V., Heeks, R., Chattapadhyay, S., & Foster, C. (2018). *Big Data and Urban Transportation in India: A Bengaluru Bus Corporation Case Study*, GDI Development Informatics Working Paper no. 72, University of Manchester, UK.

Reynolds, M. (2014). Equity-focused developmental evaluation using critical systems thinking. *Evaluation*, 20(1), 75–95.

Sengupta, R., Heeks, R., Chattapadhyay, S., & Foster, C. (2017). *Exploring Big Data for Development: An Electricity Sector Case Study from India*, GDI Development Informatics Working Paper no. 66, University of Manchester, UK.

Spratt, S., & Baker, J. (2015). *Big Data and International Development: Impacts, Scenarios and Policy Options*, Evidence Report no. 163, IDS. Falmer, UK: University of Sussex.

Taylor, L. (2017). What Is data justice? The case for connecting digital rights and freedoms globally, *Big Data & Society*, 4(2), 1–14.

Taylor, L., & Broeders, D. (2015). In the name of development: Power, profit and the datafication of the Global South. *Geoforum*, 64, 229–237.

WWW (2017). *Open Data Barometer*. Washington, DC: World Wide Web Foundation.

Young, I. M. (2006). Responsibility and global justice: A social connection model. *Social Philosophy and Policy*, 23(1), 102–130.

14

The Ethics of Big Data as a Public Good

Which Public? Whose Good?

Linnet Taylor

> "Privacy is your right. So is access to food, water, humanitarian response. The challenge is that we see a lot of regulatory frameworks which don't have the right litmus test."
>
> **Robert Kirkpatrick, head of UN Global Pulse**

> "The ethics of not acting is the ethics of being afraid of doing something wrong."
>
> **Geoffrey Canright, head of Data Analytics Group at Telenor Research**

14.1 Introduction

As data analytics have evolved over the last decades, the discussion about their uses and problems has evolved to include considerations of ethics, privacy, and legal boundaries. The notion of "data as a public good" (Olsen, Saunders, & McGinnis, 2011; Ritchie & Welpton, 2011; Tennison, 2015; UN Global Pulse, 2014) has been widely cited since the high-profile use of data from mobile phones, in particular, has begun in the humanitarian and international development sectors. The claim that it is possible for commercially generated data from mobile network operators, for example (the case explored in this chapter), to function as a public good is not self-evident. In the "data for development" challenges explored below, data is produced commercially, then channeled as a proprietary commodity into the possession of actors who may be private or public sector in origin. What, then, does this claim that "data is a public good" mean, and how should we understand its practical implications for those reflected in, and affected by the use of, such data?

Linnet Taylor, *The Ethics of Big Data as a Public Good*. In: *Human Development and the Data Revolution*.
Edited by: Sanna Ojanperä, Eduardo López, and Mark Graham, Oxford University Press. © Linnet Taylor (2025).
DOI: 10.1093/oso/9780198879145.003.0015

The two quotations that begin this chapter epitomize the perceived imperative to use the new data sources, and related to this, the need to formulate new research ethics that can guide, but also normatively promote and facilitate, their use. The value of digital data on low- and middle-income countries (LMICs) to these sectors has given rise to calls for "data philanthropy" (Kirkpatrick, 2011) where corporations donate data to nonprofit actors through development intermediaries such as the UN. There are some high-profile examples of data philanthropy, for instance where Twitter has granted access to its data to UN Global Pulse (Bellagio Big Data Workshop Participants, 2014a). However, despite their data's value to the development sector, mobile operators are proving reluctant to share Call Detail Records (CDRs) under the kind of general principles that are being discussed by development and humanitarian actors. This chapter will analyse how the argument for "data as a public good" fits with corporations' experiences as holders and owners of big data, using two case studies where mobile operators have shared CDRs for development and humanitarian purposes.

There are two possible definitions of a public good that are relevant in this scenario. One is the notion that data should be made available in order to help international organizations promote social good in the public sphere, i.e., that data should be *more public* in order to create *more good* impacts. The second is the argument that data should be formally *defined as a public good* because its potential power to fight poverty and disease, and to inform emergency response, is such that international institutions should have access to it on principle. The second is effectively an argument for reordering power relations with regard to digital data.

Defining data as a public good, however, may bring its own problems. Information economists Stiglitz (1999) and Varian (1999) have explored how information and knowledge may be considered as a public good—that is, a resource that is both nonrivalrous, with no extra cost incurred when others use it, and nonexcludable, so that it is impossible to keep others from using it. They both warn that knowledge can be made functionally excludable if the private sector gains value from controlling it, and that legal and bureaucratic regimes also determine the extent to which it is excludable, for example in the form of taxes and patents. Purtova (2015) brings this debate up to date by identifying digital data deriving from people as a "system resource" comprising an ecosystem of people, platforms, and profiles, and concluding that while it may be possible for knowledge to be a public good, it is not possible to make the same claim for digital data. In fact the language of the "personal data ecosystem" is already in use by the World Economic Forum, among others (Hamlin & Hodder, 2011; World Economic Forum, 2014), to explain the

ways in which the knowledge produced through digital data is inherently commercial and operates as an interaction between individuals and firms.

In this chapter I will build on Purtova's logic to challenge the rhetoric of data as a public good, by comparing the experiences of two firms who have shared mobile phone CDRs for research purposes: Telenor, based in Norway, and Orange, based in France. My findings are based on fieldwork, observation, and interviews conducted over the period 2012–2016, and specifically during 2015–2016 at the two firms that provide the case studies for this chapter (for a list of interviewees, please see Table 14.1.

14.2 Background: "Responsible Data" and CDRs

Since the exponential growth in mobile usership in low- and middle-income countries[1] during the 2000s (ITU 2015), CDR data from these countries have become a potentially powerful way of understanding what is being missed by national statistics, which are often lacking due to low resources (Jerven, 2013), and for planning and evaluating interventions in the development and humanitarian spheres. Large-scale datasets from international mobile operators are being used in the realm of big data analytics to model the spread of diseases such as dengue (Wesolowski et al., 2015) and cholera (Bengtsson et al., 2015), and to map poverty (Gutierrez et al., 2013) and mobility (Taylor, 2016b). Data from mobile financial transactions combined with mobile surveying is also showing potential as a way to plan savings and banking interventions (Bellagio Big Data Workshop Participants, 2014: 21), and records derived from mobile money platforms also have significant potential value for tracking income and spending.

Development and humanitarian research using CDR data began with one-off, targeted collaborations between mobile operators and researchers in public health (Bengtsson et al., 2011; Pindolia et al., 2012; Wesolowski et al., 2012), which then sparked a move to create larger-scale, more crowd-sourced forms of data sharing for research. Spain's Telefónica hosted a datathon in 2013 that opened up CDRs to the technology community,[2] and Orange established a "Data for Development" challenge[3] that released CDR data from its subsidiaries first in Côte d'Ivoire and later in Senegal to data

[1] I use the World Bank's definitions when grouping countries, see: http://data.worldbank.org/about/country-classifications, where LMICs have incomes of US$1,036—$12,616 per capita and high income countries (HICS) sit above that threshold.

[2] For more information, see http://dynamicinsights.telefonica.com/blog/1008/campus-party-2

[3] http://www.d4d.orange.com

science research teams worldwide. These two models, targeted collaboration and crowdsourcing research, are explored in the two case studies in this chapter.

The demonstrated power of CDRs to supplement national statistics and generate new insights with regard to LMICs has created a community of researchers, funders, and intermediaries interested in using them. Central to this process has been the United Nations' call for a "Data Revolution" (United Nations, 2014) and the World Economic Forum's claim that big data should be considered a tool for development and humanitarian action (World Economic Forum, 2015). Despite a growing awareness of the power of the new sources of digital data, however, no generalizable approach to data sharing has emerged. The prevailing vision is one where data is shared through intermediary institutions whose job it would be both to analyze data and to channel it directly to NGOs that could use it,[4] such as the UN data science initiative Global Pulse. Proponents of this model have put pressure on mobile operators to share data through them as intermediaries, which the firms have resisted on the basis that they remain responsible for the data they collect and should therefore determine how they are used.

The community interested in creating a public good approach to data sharing has for some time been working to create ethical guidelines to underpin such a framework, recognizing that for data about developing countries to flow more freely, rules and norms are necessary to protect their populations. These concerns were epitomized by a statement by Robert Kirkpatrick of UN Global Pulse in 2012:

> Even if you are looking at purely anonymized data on the use of mobile phones, carriers could predict your age to within in some cases plus or minus one year with over 70 percent accuracy. They can predict your gender with between 70 and 80 percent accuracy. One carrier in Indonesia told us they can tell what your religion is by how you use your phone. You can see the population moving around.[5]

It was against this background that various Responsible Data groupings and discussions emerged and evolved in two main directions after the first big-data humanitarian study in 2010 (Bengtsson et al., 2011). One was the search for models and frameworks to act as templates for long-term, organized data-sharing cooperation between corporations and the development and

[4] Robert Kirkpatrick, director, UN Global Pulse, interviewed 18 August 2014.

[5] Robert Kirkpatrick, interview with *Global Observatory*, 5/11/2012. Accessed online 7/7/2016 at http://theglobalobservatory.org/interviews/377-robert-kirkpatrick-director-of-un-global-pulse-on-the-value-of-big-data.html.

humanitarian community. A second discussion has arisen around the ethical challenges raised by the new data sources in this context.

Figure 14.1, based on fieldwork and interviews conducted during the period 2014–2016, shows the institutions and connections shaping both these discussions, as observed over the period 2013–2016. The diagram uses the Force Atlas layout algorithm, which arranges the visualization so that "brokers," or institutions with a high number of outgoing connections, are laid out towards the periphery of the graph and "authorities," or those with a high number of incoming connections, are clustered toward the centre (Khokar, 2015: 65). The network is colored according to clusters (composed of strong ties between institutions) defined using a modularity calculation (Khokar, 2015: 130–132). The names of institutions are shown proportionate to the institutions' betweenness centrality, which is the frequency with which

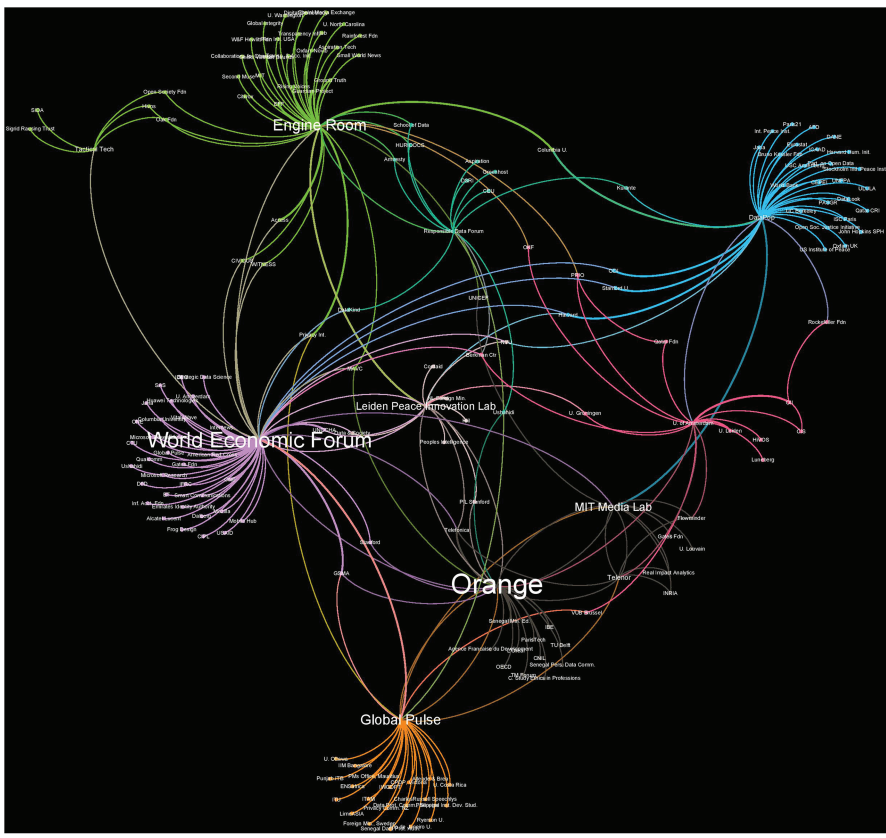


Figure 14.1 Network of participants in Responsible Data discussions, 2014–2016.

Reproduced with permission from Taylor, L. (2016). The ethics of big data as a public good: Which public? Whose good? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160126.

each is found to be the main connector between two other institutions, so that the size of the label indicates an institution's relative importance in connecting other institutions in the discussion.

The network is characterized by two key features: first, it is multipolar, with at least eight key brokerr institutions convening discussions. Second, the institutions with the highest betweenness centrality are those who stand to gain from data's increased availability to research and policymakers in terms of their funding, status, or influence.

The diagram can be interpreted as a negotiation between the larger key players over what should constitute practices of responsible data sharing, with each largely convening its own discussion community in an effort to capture the process of standardization and become the global authority. Mobile operators (Orange, Telenor, and Telefónica), academic bodies, and civil society organizations can be seen to act as both brokers and authorities. Also remarkable is the extent to which the conversation focuses on mobile data, as seen by the absence (or peripheral nature so that they were invisible to this research) of other producers of data such as governments, Internet search, and social media firms. This is in line with the fact that the conversation is largely about data from LMICs, where mobile operators are the key actors in collecting and potentially sharing data.

The high-level international actors that feature in Figure 14.1 demonstrate that there is much to play for in this debate, where funding and status potentially attach to becoming a facilitator or intermediary, and even more so to capturing the standardization process. Alliances between international bodies and academic research groups bring funding and status with them, and countries also stand to benefit: for example, both the Dutch Foreign Ministry and the Hague's city administration fund Leiden University as a convenor for data governance discussions as a way to increase the Netherlands' international profile as a promotor of humanitarian innovation.

When initial calls for data to be made more available did not result in large-scale data sharing, particularly by mobile operators, discussions toward the creation of ethical frameworks took centre stage as a strategy to present a coherent set of community standards that would make data proprietors more comfortable sharing data. This led to a clash of perspectives between the corporate owners of big data and those involved in international organizations, who perceived extreme need and strong arguments for data reuse and sharing.

In 2015 a particularly revealing example of this clash emerged when a media storm arose over the (non)use of CDRs for tracking Ebola (McDonald, 2016), demonstrating how widely they had become seen as a potential

magic bullet for emergencies and epidemics. It was implied that access was only being restricted by an over-sensitivity to privacy considerations which were far outweighed by the potential good of the data (Economist, 2014; Talbot, 2014). In reality, however, the applied research done using CDR data has been incremental and long-term, involving overlapping scientific, humanitarian, and core business concerns.

The discussions around sharing CDRs demonstrate Purtova's ecosystem dynamic. To frame data as a public good, the broker institutions need agreement from the whole ecosystem on appropriate standards. Yet the incentives conflict: broker institutions are motivated to standardize because it will enable them to serve as brokers, but the authorities gain from the discussions rather than from standardization itself. Mobile operators play a dual role as brokers who benefit from research collaborations and as authorities who wish to determine how those collaborations should function. The following section will explore the considerations at play in developing research collaborations, and in particular the role privacy and ethics play in shaping them.

14.3 Choosing Collaborations: The Influence of the Ecosystem

At the research department of the Norwegian mobile operator Telenor in late 2015, studies were being conducted with outside collaborators using CDRs to detect poverty levels in low- and middle-income countries, and the firm had just released a study conducted in Pakistan that showed CDRs could predict the transmission of dengue (Wesolowski et al., 2015). At the same time in Paris, Orange was funding follow-up work from its "Data for Development" challenge and working on its next approach to sharing data on developing countries: an "open algorithm" project (OPAL), that would act as a bridge into the data held by the company and allow researchers to ask questions of the data without the firm having to release it to them.

Sharing CDR data is always a risk. In its raw form the data is highly privacy sensitive, and sharing it, even in de-identified form, means legal checks to comply with data protection laws, precautions to make sure competitors do not gain access to privileged information, and input from CSR specialists to make sure the exercise will not result in bad publicity. Moreover, when de-identified CDR data is shared, the mobile operator gives consent for the data to be reused, not the firm's customers. This puts an extra onus on the firm to

be cautious, because it is taking individuals' privacy into its own hands. On balance, it is safer not to share data. However, CDRs are increasingly shared for research. Why, then, does this occur?

14.4 Business Advantage

According to interviewees for this project, CDRs are shared in a situation where the firm perceives an overall benefit from sharing them, both in terms of business advantage and social impact (which also, ideally, confers a business advantage over the longer term). Pål Sundsøy, a Senior Data Scientist at Telenor Research, described the company's research into dengue transmission in Pakistan as being situated precisely at this intersection:

> In the long run it seems important to have good media impact. You choose projects based partly on business value if you want to get them approved . . .
>
> Dengue is pretty altruistic in terms of our projects, we want to be a positive force in the markets where we operate.

Researcher Taimur Qureshi of Telenor similarly stated that he saw the process of using data as being "a positive force in the markets where we operate." This project clearly did achieve media impact: a week after the study's publication more than 100 positive articles had appeared online around the world. Geoff Canright, head of the Data Analytics Group at Telenor Research, explained that the choice of dengue (a disease that affected people across geography and socioeconomic status) was influenced by the desire not to create tensions with the Pakistani government by creating politically sensitive datasets. Job performance criteria were similarly a motivation for researchers at Telenor to create projects at this intersection of business and altruism: employees are judged on both their internal and external value creation, and scientific publications and advisory work are named as important components of good performance.

At Orange, the reasons to facilitate CDR research with their Data for Development challenge and subsequently with the OPAL project were similarly a mixture of altruism and core business. The first Data for Development challenge was at least partly driven by curiosity: an effort to go beyond the established model of collaborating with outside researchers to research business challenges such as predicting churn.[6] In each case of data sharing,

[6] Interview with Prof. Vincent Blondel (March 29, 2013).

altruism alone is not sufficient for a business case, but a perception of social value is necessary to overcome the business risks of sharing data outside the company.

14.5 Data-Sharing Models and Underlying Strategies

Beyond the social aspects of data sharing—perceptions of advantage and one-time opportunities—data sharing also depends on firms being able to create the right technical model for channeling and sharing the data. De Montjoye et al. (2016) identify four main models currently in use: remote access, limited release, a question-and-answer model, and the use of pre-computed indicators and synthetic data. Of these four models, Orange's D4D challenge took the form of a limited release of data: researchers received an initial description of the data, then sent in abstracts on the basis of which they received four datasets. The datasets showed traffic between antennae, movement trajectories for a large group of phone users referenced by antenna location, movement trajectories according to country administrative regions, and communication network details for a group of users. The data were de-identified by Orange's local subsidiary, and antennae locations were blurred in order "to protect Orange's commercial interests" (Blondel et al., 2012). This model for data sharing allowed researchers to map and track mobility as well as communication network activity in terms of space and time, but without knowing the names of the users in question.

Telenor's collaboration with Harvard Public Health was more conservative in that the data was not only de-identified by hashing and encrypting, but was also aggregated and processed by the local subsidiary (Buckee, 2015). Following the remote access model, the Harvard researchers worked with Telenor's research department to identify the maximum spatial and temporal aggregation level for the data that would still make it possible to answer the research question. The Telenor researchers aggregated the data to base-station level locally at the Pakistan affiliate's offices, making it possible to move the data out of the country. At the firm's Oslo headquarters they then aggregated the data up another level due to concerns about business sensitivity, so that it showed movements of mobile phones not between base stations but between Tehsils, Pakistani administrative units below the province level. The Harvard team were then given access to the aggregated matrices, rather than the dataset itself as with the Orange challenge.

In the case of Orange, an unexpected problem occurred once the data had been shared and the research outputs made public. The winning paper from

the first D4D challenge (Berlingerio et al., 2013) was written by a team based at IBM. This created a business conflict, since in order to follow up research with action, as Orange had originally hoped, further data and funding would have to be shared with a competing business. This problem led to an imaginative response by Orange: the OPAL (Open Algorithm) project, which aims to provide an open platform that will shield the raw data in the firm's CDRs while allowing outside researchers to extract answers from them (Orange, 2016).

The complexity of the arrangements involved in collaborations demonstrates why mobile operators are reluctant to endorse any standardized procedure designed to facilitate the operability of data as a public good. There is no single level of aggregation or de-identification strategy that will render data safe for all users. This is the problem Orange's OPAL solution is designed to remedy: operators gain from the use of their data by outside researchers, but only as long as they keep control of their data. Even this, however, does not fully address the ecosystemic nature of digital data stemming from human activities, and therefore requires more. The question of what constitutes ethical use from the point of view of the operator will be considered in the next section.

14.6 Ethical Questions

Data protection laws in the USA and EU protect raw CDR data from being shared for reasons of privacy. However, given the potentially greater sensitivities of data about LMICs (Taylor, 2016b), law is not the only issue to consider in sharing data. When the purpose of sharing data is to identify poor, vulnerable, or disadvantaged groups, even the most cautious data-sharing models may only create more nuanced problems of consent and potential harm than data protection laws currently envisage.

14.7 Consent

Geoffrey Canright of Telenor identifies consent as central to the process of sharing data. "Informed consent is understanding what you are saying yes to, and knowing what data are being used and why. You should have the option to take your consent back at any time." However, in the case of mobile operators sharing data with outside researchers, Nikolai Pfeiffer, Telenor Group Privacy Officer, noted that "there are two layers: consumer

consent and affiliate consent." The mobile phone user consents via their original contract with the operator to their de-identified data potentially being used for research purposes. Then the affiliate operator must consent to the firm itself using the data. The same process was described by executives at Orange. This raises the potential problem that informed consent to the use of data in research on the part of customers is very general and does not involve revisiting the contract when data is repurposed. In the case of Telenor, the company addresses this by making informed consent on the part of the country affiliate as specific as possible to a single use of data, and checking back if new uses are proposed. In this case, the weakness is that the affiliate takes the place of the individual in the consent process, but the strength is that purpose limitation is observed.

14.8 Identifiability

The second question that arises from these processes of data sharing is that of pinning down the concept of identifiability. What is personal information in the case of de-identified calling data? This is not legally a problem as long as data is de-identified. Canright explains that working internationally means dealing with multiple layers of regulation:

> We have to invent procedures, the regulators look at Personal Information . . .
>
> Our data is deidentified. If you aggregate location data over weeks, no way is that personal data—unless everyone goes to the same place at the same time as a group.

The group problem, however, is possibly the next frontier for privacy. As big data began to be used to track disease, conflict, and population movements in ways (anonymized and aggregated) that were conventionally considered safe according to data protection norms, it became clear that these data nevertheless posed a risk to those who generated them. For example, if an adversary is interested in harming people of a particular ethnicity or religion, individual identities are relatively unimportant. Instead, the adversary will want to know which villages, networks, or gatherings to target. Thus while data protection norms and the corresponding rules and guidelines focus on safeguarding individual identity, it is possible to envisage situations where even de-identified data may be sensitive. Big data such as CDRs allow people to be tracked as groups and networks (Taylor, 2016a, 2016b), so that

data aggregated for disease prediction can instead be used to track groups of political interest such as separatists, smugglers, undocumented migrants, or dissidents. A number of institutions working in the field of development and humanitarian data arrived independently at the group privacy problem over the period 2010–2016 (Orange, 2016; Raymond, 2016; UNOCHA, 2014), as did scholars working on the challenges of new data technologies (Floridi, 2014; Raymond, 2016; Taylor, 2016b).

The dengue project was almost an ideal case from an ethical point of view because the results of the model could be shared with public health authorities in the form of paper maps rather than sharing the data or the model itself. Tracking many other diseases, however, requires disaggregated data. In the case of Ebola (McDonald, 2016), it is necessary to know not only where people are traveling but what they are doing and what kind of contact they have with each other; something McDonald points out is so far beyond the law that it would require government to commandeer raw data from mobile operators.

Taimur Qureshi, senior data scientist in Telenor's research group, explained how sensitive location data can be:

> Even de-identified but nonaggregated data can be sensitive because you can identify someone because of his location. Location patterns also reflect behaviour. [The data shows] if you are up at night, if you are male or female—men have a larger radius of gyration away from home—your salary, because those who earn more tend to move more widely and often, and to socialize more widely.

This level of sensitivity in even de-identified data suggests that the "right" way to share and use CDR data is a moving target that must be addressed with social and political knowledge of the data's original context, and that depends on whom the data is shared with, and in what form.

14.9 Conclusion: Can CDR Data be Framed as a Public Good?

This chapter has argued that to understand why the notion of data as a public good has not gained traction with its target audience, it is necessary to see from the point of view of that audience. Rather than censoring out of sheer risk aversion, corporations are navigating the challenges of CDRs' potential for both good and bad in nuanced, reflexive ways that result in the evolution of new ethical approaches to data-sharing infrastructures

and practices. They also experience data production as an ecosystem where different stakeholders have contractual rights, interests, and property claims. On this basis, firms are highly incentivized to keep control of data from LMICs because they have high value as a proprietary resource that contains insights about business processes and customer preferences as well as the potential for professional learning and positive perceptions on the part of customers and governments—and this value is balanced by a high negative potential for the firm if the data are misused.

One underlying reason for certain actors not wanting CDRs to become a public good is that this value is linked to their scarcity. The calls for data to operate as a public good effectively highlight the fact that it does not currently do so, and the debate highlighting CDRs' importance to researchers and policymakers effectively increases the status of firms as gatekeepers of high-value data. Out of the two overarching models for data sharing—one where an intermediary holds and provides access to multiple datasets, versus another where firms control the process of data-sharing—the second is likely to be more attractive to data proprietors because it both creates scarcity and focuses the good publicity arising from such projects mainly on the firm.

The ecosystemic characteristics of private-sector data demonstrated in this chapter make it harder for data to be shared without an idea of its eventual purpose. Firms experience data-sharing projects as single-use constructs created through negotiations with a constellation of different actors including country affiliates, country governments and security services, the firm's customers, and different departments within the firm's headquarters including CSR management and privacy officers. Furthermore, research collaborators' domain knowledge also defines how data must be channeled and packaged to address a particular research problem. This ecosystem makes data use and sharing highly contingent, contextual, and incremental, and highlights the functional reasons why firms are not sharing data more freely.

All claims that data should be a public good rest on an (often unspoken) assumption that procedures for preventing harm can be standardized. This chapter has demonstrated why this is a difficult assumption to make even in the case of a single type of data and a single genre of use. Significantly, the claim that an ethical approach to data sharing can, and must, be standardized has been made most urgently and received most institutional support in the field of development and humanitarian response. These fields, however, must employ great caution about understanding the ecosystem in which they are operating. Orange's OPAL project can be seen as an attempt to sidestep this ecosystem problem by ensuring the dataset and its value remain within the firm's auspices: companies will use the OPAL platform to allow others

to ask questions of their proprietary databases without actually releasing the data, and the algorithms can be regulated, controlled and certified by official bodies based on ethical considerations. However, as with the other models this does not address the question of who has ultimate ethical responsibility for the use of the research outputs. The OPAL approach avoids the utilitarian cost–benefit analysis by placing the onus on public certification bodies to make sure that algorithms and their output are not being used in harmful ways. This may be the system's chief benefit in terms of data ethics: by removing layers of consent and permission it draws attention to the unpredictable outputs of sharing CDRs, and the necessity for a separate layer of project vetting by public bodies.

OPAL would effectively allow proprietors to rent out their datasets, and would shift ethical responsibility for the use of the data to a public body (at least as the system is currently conceived). One potential benefit of OPAL would be that, like the current system of limited sharing, it imposes the principle of purpose limitation on data analysts, forcing researchers to approach the data with an already defined problem rather than mining the data for all possible correlations. This in turn limits the possibility that researchers with little contextual knowledge will be able to mine potentially sensitive data about LMIC populations, thus answering one central ethical problem with regard to the practice of data science for development (Taylor, 2016b).

The two quotations that began this chapter are fundamentally mutually compatible. It should be possible to both claim that LMIC data should be used for the benefit of all, and for corporate data proprietors to be active in providing those data. However, the Responsible Data debate has not been able to overcome the obstacles to this by producing technical and ethical frameworks within which data can operate as a public good (see, for example, Antin et al. 2014). It has, however, contributed significantly to understandings of data ethics amongst a highly diverse disciplinary and professional community by demonstrating that the search for a litmus test, a single set of guidelines or a decision-making template will ultimately fail because it asks the wrong question. By asking how to make datasets safe for use through the "right" processes of de-identification or aggregation, one is effectively asking the notion of "data ethics" to carry the twofold burden of conceptualizing the right to privacy and autonomy across widely varying cultural and political contexts, and of substituting for legal parameters where these are lacking.

The claim that data should be a public good is important to evaluate because it is a statement about power, both power over data and power over the outputs data can produce. The struggle to find an overarching framework

for data sharing and use is also the struggle to assert a claim over data on the basis of moral necessity. Creating greater access to corporate datasets would distribute some of the power currently held by firms to intermediaries and researchers, and, by incorporating more decision-makers and less purpose limitation in data's value chain, would also shift power away from the individuals who originally generated the data and who are most likely to be affected by possible misuse. Since an overarching framework, as currently conceptualized, would effectively substitute for legal or regulatory measures in LMIC contexts where those measures do not yet exist, corporate resistance to this strategy may have positive effects on the development of the new field of data ethics. This resistance highlights an institutional gap in terms of independent public bodies capable of auditing and authorizing new uses of corporate data. Without independent oversight, it is unrealistic to expect LMIC data to flow freely amongst guardians who propose to guard themselves.

Acknowledgments

The research for this chapter was conducted as a Marie Curie postdoctoral fellowship at the University of Amsterdam. I am grateful to Geoff Canright of Telenor and both Nicolas de Cordes and Zbigniew Smoreda of Orange for granting me access to their research groups for interviews and observation, and to Scott Hale and Taha Yasseri for advice on network visualization.

Appendix 1

Table 14.1 Interviewees contributing to the research.

Name	Title	Institution	Date interviewed
Vincent Blondel	Professor, Applied Mathematics	Université Catholique de Louvain	29 March 2013
Geoffrey Canright	Head of Data Analytics Group	Telenor	17 September 2015
Robert Kirkpatrick	Director	UN Global Pulse	18 August 2014
Nikolai Pfeiffer	Telenor Group Privacy Officer	Telenor	16 September 2015
Taimur Qureshi	Senior Data Scientist	Telenor	16 September 2015
Pål Sundsøy	Senior Data Scientist	Telenor	16 September 2015

References

Antin, K., Byrne, R., Geber, T., van Geffen, S., Hoffmann, J., Jayaram, M., . . . Wilson, C. (2014). Shooting our hard drive into space and other ways to practice responsible development data. *Online Publication*. Available at: https://archive.ids.ac.uk/eldis/document/A69542.html

Bellagio Big Data Workshop Participants. (2014). *Big data and positive social change in the developing world: A white paper for practitioners and researchers*. Oxford. Retrieved from http://www.rockefellerfoundation.org/uploads/files/c220f1f3-2e9a-4fc6-be6c-45d42849b897-big-data-and.pdf

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., . . . Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports*, *5*, 8923. http://doi.org/10.1038/srep08923

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *PLoS Medicine*, *8*(8), e10010831–9. http://doi.org/10.1371/journal.pmed.1001083

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., & Sbodio, M. L. (2013). All Aboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. In H. Blockeel, K. Kersting, S. Nijssen, & F. Železný (Eds), *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013*, *Prague, Czech Republic*, *September 23–27*, *2013*, *Proceedings, Part III* (pp. 663–666). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-40994-3_50

Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., . . . Ziemlicki, C. (2012). Data for development: the D4D challenge on mobile phone data, 1–10. Retrieved from http://arxiv.org/abs/1210.0137

Buckee, C. O. (2015). Mobile phone data for public health: A data-sharing solution to protect individual privacy and national security.

De Montjoye, Y.-A., Blondel, V., Canright, G., Gambs, S., Garcia, M., Kendall, J., . . . Smoreda, Z. (2016). Privacy-conscientious use of mobile phone data. *MIT Working Paper*. Retrieved from http://openscholar.mit.edu/sites/default/files/bigdataworkshops/files/draft_modelsformobilephonedatasharing.pdf

Economist (2014). Waiting on hold. *Economist*. Retrieved from http://www.economist.com/news/science-and-technology/21627557-mobile-phone-records-would-help-combat-ebola-epidemic-getting-look.

Floridi, L. (2014). Open data, data protection, and group privacy. *Philosophy & Technology*, *27*(1), 1. http://doi.org/10.1007/s13347-014-0157-8

Gutierrez, T., Krings, G., & Blondel, V. D. (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv:1309.4496 [Physics]*, 6. Retrieved from http://arxiv.org/abs/1309.4496

Hamlin, K., & Hodder, M. (2011). Personal data ecosystem response to FTC Do Not Track Proposal, 18/2/11. unpublished. Retrieved from https://www.ftc.gov/sites/default/files/documents/public_comments/preliminary-ftc-staff-report-protecting-consumer-privacy-era-rapid-change-proposed-framework/00472-57974.pdf

Jerven, M. (2013). *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It*. Ithaca: Cornell University Press.

Khokar, D. (2015). *Gephi Cookbook*. Birmingham: Packt.

Kirkpatrick, R. (2011). Data philanthropy: Public and private sector data sharing for global resilience. Retrieved from http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience

McDonald, S. M. (2016). *Ebola: A Big Data Disaster Privacy, Property, and the Law of Disaster Experimentation* (March 2016 No. 2016.01). Bangalore. Retrieved from http://cis-india.org/papers/ebola-a-big-data-disaster

Olsen, L., Saunders, R. S., & McGinnis, J. M. (2011). *Clinical Data as a Public Good for Discovery*. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK92063/?report=reader

Orange. (2016). OPAL. Retrieved from http://www.data4sdgs.org/dc-opal/

Pindolia, D. K., Garcia, A. J., Wesolowski, A., Smith, D. L., Buckee, C. O., Noor, A. M., … Tatem, A. J. (2012). Human movement data for malaria control and elimination strategic planning. *Malaria Journal*, *11*(1), 205. http://doi.org/10.1186/1475-2875-11-205

Purtova, N. (2015). The illusion of personal data as no one's property. *Law, Innovation and Technology*, *7*(1), 83–111. Retrieved from file://unimaas.nl/users/fasos/tamar.sharon/data/MyDocuments/Downloads/SSRN-id2762225.pdf

Raymond, N. A. (2016). Beyond "Do No Harm" and Individual Consent: Reckoning with the Emerging Ethical Challenges of Civil Society's Use of Data. In L. Taylor, L. Floridi, & B. van der Sloot (Eds), *Group Privacy: New Challenges of Data Technologies*. Springer.

Ritchie, F., & Welpton, R. (2011). *Sharing risks, sharing benefits: Data as a public good*. *In*: *Work session on statistical data confidentiality, Eurostat, Tarragona, Spain, October 26–28, 2011*. Retrieved from http://eprints.uwe.ac.uk/22460

Stiglitz, J. E. (1999). Knowledge as a global public good. In *Global Public Goods: International Cooperation in the 21st Century* (pp. 308–325). Eds: Kaul, I.; Grunberg, I.; and Stern, M.A. New York: Oxford University Press, http://doi.org/10.1093/0195130529.001.0001

Talbot, D. (2014). Cell-Phone Data Might Help Predict Ebola's Spread. *MIT Technology Review*. Retrieved from http://www.technologyreview.com/news/530296/cell-phone-data-might-help-predict-ebolas-spread

Taylor, L. (2016a). Data Subjects or Data Citizens? Addressing the Global Regulatory Challenge of Big Data. In *Freedom and Property of Information: The Philosophy of Law Meets the Philosophy of Technology*. (pp. 81-105). Eds: Hildebrandt, M.; van den Berg, B. London: Routledge.

Taylor, L. (2016b). No place to hide? The ethics and analytics of tracking mobility using mobile phone data. *Environment and Planning D: Society and Space*, *34*(2), 319–336. http://doi.org/10.1177/0263775815608851

Taylor, L., & Broeders, D. (2015). In the name of development: Power, profit and the datafication of the Global South. *Geoforum*, *64*(4), 229–237. http://doi.org/10.1016/j.geoforum.2015.07.002

Tennison, J. (2015). Why is open data a public good? Retrieved from https://theodi.org/blog/why-is-open-data-a-public-good

UN Global Pulse (2014). *Annual Report 2014*. New York. Retrieved from http://www.unglobalpulse.org/sites/default/files/Annual-Report_2014_FINAL-DIGITAL-FOR-EMAIL_0.pdf

United Nations (2014). *A World that Counts: Mobilising the Data Revolution for Sustainable Development*. New York: United Nations.

UNOCHA (2014). *Humanitarianism in the Age of Cyber-warfare: Towards the Principled and Secure Use of Information in Humanitarian Emergencies* (OCHA policy and studies series No. 11).

Varian, H. R. (1999). *Markets for Information Goods. Institute for Monetary and Economic Studies, Bank of Japan* (Vol. 99). Retrieved from http://people.ischool.berkeley.edu/~hal/Papers/japan/index.html

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, 267–270. http://doi.org/10.1126/science.1223467

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. a., Rasheed, S. B., . . . Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, 201504964. http://doi.org/10.1073/pnas.1504964112

World Economic Forum. (2014). *Rethinking Personal Data: Trust and Context in User-Centred Data Ecosystems. Rethinking Personal Data: Trust and Context in User-Centred Data Ecosystems*. Geneva. Retrieved from http://www.weforum.org/reports/personal-data-emergence-new-asset-class.

World Economic Forum. (2015). *Data-Driven Development Pathways for Progress. Data-Driven Development Pathways for Progress*. Geneva.

Index

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.

Tables and figures are indicated by t and f following the page number.

"4 Cs" of Big Data, AI and the Data Revolution 94–97
  see also three functions of data
2030 Agenda for Sustainable Development 3, 38–39
  data funding gap 38–39

academic publishing 10–11
Addis Ababa 80–81
administrative data 216–217
Africa Agenda 2063 71–72, 80
African Data Consensus (2015) 72, 80
African Union Data Policy Framework 47–48
Africa's "statistical tragedy" 118
AI *see* Artificial Intelligence
Akoya.com 141
Alalawneh, A. 8
algorithmic fairness 44, 140
algorithmic transparency 66, 138–139
algorithms
  algorithmic bias 206–207, 243–244
  personalization algorithms 51–52
  "tyranny of algorithms" 139
  see also data-driven algorithmic policy-making; OPAL project
Alkhatib, S. 8
application programming interfaces (APIs) 241
archiving of data 191, 197
Artificial Intelligence (AI) 90, 238
  "Human AI" 102
  see also "4 Cs" of Big Data, AI and the Data Revolution; machine learning; natural language processing; Uganda
Audio Computer Assisted Self-Interviewing (ACASI) 215–216
Automatic Speech Recognition (ASR) 162, 164, 168

"better data, better decisions, better lives" 91, 101
bias 77, 102, 103, 131, 151, 168, 207, 243
big data 118–119
  big data analysis 4–5, 8, 12, 14, 29, 37–39, 41
  big data sources 16–17, 32, 35–36, 43
  big data for development 13–14, 154, 205
  big development data 43, 187–188
  combining big data with other data 119
  definitions of 6–7, 94
  evolution of data from "small" to "big" 231–233
  limitations of 119, 206–207, 217
  sources of big data *see* sources of big data
big data as a mindset 205–206
"Big Data" ecosystem 6–7
Big Data for Social Good 9, 52–54, 65
  access to real-time data 67
  challenges 65–67
  financial challenges 65
  privacy and security challenges 65
  regulatory challenges 65
  social challenges 65
  technical challenges 65
  uses of 53
  way of working challenges 65
  see also crime; flooding in Tabasco; natural disasters
"big data paradigm" 6–7, 35–36
Blumenstock, J. 120
Brazil 79
Bronfenbrenner, U. 78, 85

call detail records (CDRs) 119–120, 141–142, 149
  combining with other publicly available datasets 120

estimates of socioeconomic indicators 120
studies using 121
see also Rwanda; sharing CDR data for research purposes
Canright, G. 280
Capacity Development 154 see "Statistical capacity"
causality 207, 232–233
CDRs *see* call detail records
census data
poverty indices 118
challenges and barriers relating to data-driven development 8, 13, 30, 34, 43–44
differing capabilities, remits, and resources among development actors 36–37, 44–46
financial resources required to integrate big data analysis 38–39, 46–47
regulatory and ethical considerations 39–40, 47–48
representativeness, appropriateness, and quality of big data sources 35–36, 44
securing political commitment to integrate and use big data 37–38, 46
climate change 55
cognitive bias 102–103
computational research methods 4–5, 7–8
computational skills 44–46, 213–214
computational social science/computational social sciences 7, 51–52, 90
conferences and events focused on big data and data science 12
cost savings 33–34, 42–43
COVID-19 pandemic 20
crime 58–59
economic impact of 54–55
use of big mobile data to determine crime hotspots 59–65
critical data studies 40, 265–266
contractual agreement 188
data 1–2
data for development 10, 270, 277–278
data governance 181–182
data handling/data handling process 190, 196–198
data characteristics 188, 192–195, 198–203
data storage 195–196
data use 188, 194–196, 198–203
data justice 39–40, 47–48, 253–265
data life cycle/data handling life cycle 188, 190–191
data acquisition 188, 190–191, 195
data creation 188–189, 202, 226
data destruction 188, 191, 202
data management 188, 191, 197
data partnerships 10, 152–153
data quality 190, 262–263
data readiness 173
data representativeness 35–36, 44, 65, 206–207
data revolution 2–3, 90, 92–97, 137, 205–207, 273
data sampling 217
data skills 8, 36, 44–46, 153–156
data-driven policy making 138–140
legal requirements 193–194
cultural identity in Africa 85–86

data acquisition 190–191, 196
data as a public good 270, 281–284
defining data as a public good 271–272
definitions of a public good 271
ethical issues 273, 275–276, 282–284
Responsible Data discussions 273–276, 283
see also sharing CDR data for research purposes
databases 234, 241
data-driven algorithmic policy-making 139–140, 142
"tyranny of algorithms" 139
see also OPAL project
data engineering 188, 241–242
datafication 151, 261–263
"data gaps" 90–91
data handling 187–189, 202, 203
case study 198–203
data assessment 196
data characteristics 188, 192–194
data engineering 188
heuristic framework 196–198
information security 188
legal considerations 193–194
matrix 198
use and users of data 188, 194
data handling life cycle 188, 190–191
data acquisition 190–191, 196

data handling life cycle (*Continued*)
data management 191, 197
disposal/archiving of data 191, 197
data handling process 190, 195–196
data-intensive development 17–18, 31–32, 253, 267–268
data justice 39–40, 253
mainstream perspectives on 253
see also structural data justice
Data-Justice-for-Development Manifesto 267–268
data management 191, 197
data management plans 188–189
data philanthropy 35–36, 271
Data-Pop Alliance 54, 141–142
data revolution 90, 92, 94, 118–119, 136–137
mixed record of 96
New Data Revolution 93P1
potential to contribute to sustainable human development 96–97
see also "4 Cs" of Big Data, AI and the Data Revolution
data science
for governance 173–174
human capital pipeline 182
research methods 182
skills 177
Data Science for Insight and Decision Enablement (DSIDE) 177–178
challenges and wins 178–180
data scraping 241–242
data sharing 53–54, 67
see also data as a public good; exclusionary data-sharing practices
Datathon for Social Good 59
de Soto, H. 83
decentralized systems 228
development
data-driven development 31–38, 44, 47–48
challenges of data-driven development 29–48
barriers of data-driven development 30
opportunities of data-driven development 29–48
digital development 9–10
economic development 9, 58, 78
global development 2–3, 5–12, 20, 36–37
human development 4, 14, 91, 158, 159
human progress 97–98, 102, 104
socioeconomic development 15–16
development work 9–10
harms and risks of data 10
key development cooperation partners 10
themes related to data availability 9–10
differing capabilities, remits, and resources among development actors 36–37, 44–46
digital divide 1–2, 10, 75, 159
digitally enabled survey methods 215–216
digital mapping 81
Kibera, Kenya 81–85
digital platforms and social media 213–214
Digital Revolution 1–2
digitization 5–6, 51, 262–263
disposal of data 191, 197
DSIDE 177–180
Dunning-Kruger effect 102–103

Eagle, N. 120
earth observation data 120, 123, 211–212
example applications 211–212
Easterly, W. 139
ecological systems theory 74, 81
economic impact of big data 52
"end of theory" 233
ethical and regulatory considerations 39–40, 47–48, 193–194
see also data as a public good
ethics 47–48, 270–284
European Data Economy 52
exclusionary data-sharing practices 74–75
individual level 75
institutional exclusions 76
extreme poverty 117

financial resources required to integrate big data analysis 38–39, 46–47
flooding in Tabasco (Mexico)
use of big mobile data to develop early warning mechanisms and to support decision-making processes 56–58
Flowminder Foundation 54
Fourth Industrial Revolution 1–2

generativity 262–263

geographical coverage of development data, increasing 32
global big data and business analytics (BDA) market 52
global development 8, 9
Global Partnership on Sustainable Development Data (GPSDD) 10, 137
Global South 8, 146–156
Gross Domestic Product (GDP) 97–99, 105

HDI *see* Human Development Index
Heeks, R. 253–256
"Human AI" 102
  see also "learning loops"
Human Development Index (HDI) 98
  correlation between SCI and HDI 99–101
  Uganda 158
human ecology theory 72–74
human rights 86

Independent Expert Advisory Group (IEAG) 10, 76–77
India
  Bengaluru Metropolitan Transport Corporation (BMTC) 256, 261
  electricity corporation 256, 258–259, 261
  marginalized urban communities 256, 261
  Transparent Chennai 256
information security 188
information value chain 253, 255
integrated national data systems 36–37
International Development Research Center (IDRC)
  Big Data for Development Network 155
international organizations
  publications 11
Internet 1, 74–75, 213–214
Internet of Things (IoT) 51, 212–213

Jacobs, J. 58–59

Kenya
  digital mapping Kibera 81–85
  open-data initiatives 78–79
  Kenya Open Data Initiative (KODI) 79
  Kenya National Bureau of Statistics 83–85
  Kibera 81
Kirkpatrick, R. 273
knowledge transfer 10–12, 40–41

Landscan population database 123–125
language, data as 106–107
Large Language Models (LLMs) 90, 239
large-scale data *see* big data
"learning loops" 44–46, 102–105
  data as language 106–107
  data as lens 106
  data as lever 107
  three functions of data 105–107
lens, data as 106
lever, data as 107
LIRNEasia 146–147
  access to data 152–153
  advisory group 154
  combining big data research with more traditional data 147–149
  human resource issues 153–154
  influencing policy 152
  research using big data 147–150
  socioeconomic implications of using big data 150–151
  strategic engagement and catalysis 155–156
  Sustainable Development Goals 155
  way forward for 156
LLMs 90, 239
low- and middle-income countries (LMICs) 8, 37, 39, 47, 271–273, 275, 279, 282

machine learning (ML) 232–233, 237–238
  determinants of success for government projects 173
  see also South Africa
MDGs 72, 92, 99
measurement 33, 85, 90–91, 98–99, 102, 105–109
measurement-development nexus 91–92, 101–102, 108–109
  correlation between SCI and HDI 99–101
  historical analysis 97–99
  measurements as socio-political statements 99
  OPAL approach 108
  technological implications and social innovations 108

measurement-development nexus (*Continued*)
see also "Human AI"; "learning loops"
methodological approaches to the study of data 234–236
academic/practitioner perspectives 242–243
application programming interfaces 241
Artificial Intelligence 238
choice of 244
classification of information 234–235
databases 241
data engineering 242
data scraping 241–242
evolution of data from "small" to "big" 231–233
Large Language Models 239
limitations 243–244
machine learning 237–238
modeling techniques 242
natural language processing 238–239
network science 239–240
spatial statistical methods 240
time series analysis 237
traditional methodological approaches 226–228, 231–232, 235–237
Mexico 14–15, 42, 56, 79
microdata 94–96, 216–217
microfinance data 214–215
Millennium Development Goals (MDGs) 38, 72, 92, 99
ML *see* machine learning
mobile money 210
mobile network data 59
mobile phones 44, 119, 125–126, 159, 209–211, 279–280
mobile phone data 209–211
example applications 210
see also Big Data for Social Good; call detail records
mobility data 41–42, 209–211, 217
modeling techniques 242
monetary poverty indices 117
Moore's law 5–6
multi-dimensional poverty index (MPI) 122–123
see also Rwanda

Nairobi 75, 81–83
natural disasters 54–55
big mobile data 55
developing countries vulnerability to 55
economic impact of 55
see also flooding in Tabasco
natural language processing (NLP) 238–239
network science 239–240
Newman, O. 58–59, 64
Non-governmental organizations (NGOs) 72, 77, 120, 136–137
nightlight 120, 211
non-monetary poverty indices 117
Norway 79, 272

Obama, B. 79
OECD 3, 37–38
OPAL (Open Algorithms) project 108, 138, 140–142
new social contract, as 138
Open Government Partnership (OGP) 79
Brazil OGP Conference 79–80
operational and strategic insights 32, 41–42
opportunities and potentials of big data analysis for development work 12, 29, 31–34, 41
cost savings 33–34, 42–43
increased temporal detail 42
increasing the geographical coverage of development data 32
operational and strategic insights 32, 41–42
real-time and near-real time observation 33
Orange 53–54, 272–273, 276, 277–279, 282–283

participant–observer methodologies 77
personal data ecosystem 271–272
personalization algorithms 51–52
Philippines 79
policy influence and change 118, 172
poverty indices 117–118
census data 118
monetary indices 117
multi-dimensional poverty index 122–123
non-monetary indices 117
survey data 118
predictive analytics 53, 59, 118
privacy 47–48, 65, 119

probabilistic setting 228–231
probabilities 229, 231, 233, 237
public good *see* data as a public good
Pulse Lab Kampala 161, 163, 166, 169
Purtova, N. 271–272, 276
Python 228

Qureshi, T. 281

radio 41–42, 160–162
real-time and near-real time observation 33
regulatory and ethical considerations 39–40, 47–48, 193–194
  see also data as a public good
remote sensing 33, 118, 172 and 211–212
  see also earth observation data
Renken, J. 253–256
representativeness, appropriateness, and quality of big data sources 35–36, 44, 151, 168
Richards, E. S. 73–74
Rwanda
  CDR data, other publicly available datasets, and socioeconomic status of individuals 120, 132–133
    average daily call volume per phone 125, 127, 128
    datasets used 122–125
    feature engineering 125–126
    future work 131–132
    limitations of the study 131
    methodology 122
    mobile ownership per capita 125–128
    nightlight intensity per capita 125, 127, 128
    population density 126–128
    results 126–130
  censuses and surveys 118

sample space 229
satellite 120, 123, 211–212
satellite data 42, 44, 54, 92, 120, 146–147, 150, 208–209, 211–212, 226
  see also earth observation data
search engines 214
SDGs *see* Sustainable Development Goals
securing political commitment to integrate and use big data 37–38, 46
Sen, A. 254
Senegal
  correlation between SCI and HDI 101
sensitive personal data 193–194
sensitivity of data 192–193
sensor data 212–213
  example applications 212–213
sharing CDR data for research purposes 272–273, 276–281, 284
  business advantage and social impact 277–278
  consent 279–280
  ethical issues 279–281
  identifiability 280–281
  Orange 272–273, 276, 277–279, 282–283
  technical models and strategies 278–279
  Telefónica 272–273
  Telenor 276–280
social contract
  OPAL project as 138
Social Good 53
social media and digital platforms 213–214
socioeconomic 120, 132, 149
sources of big data 207–209, 217
  administrative data 216–217
  digitally enabled survey methods 215–216
  digital platforms and social media 213–214
  earth observation data 211–212
  microfinance data 214–215
  mobile phone data 209–211
  sensor data 212–213
South Africa
  Council for Scientific and Industrial Research (CSIR) 44–46, 173–174, 177–178
  data accessibility and governance 181–182
  data science 174
  Data Science for Insight and Decision Enablement (DSIDE) 177–178
    challenges and wins 178–180
  Deep Learning Indaba 177, 180–181
  energy management for municipalities 174–175
    project scoping and outcome monitoring 175–177
  Square Kilometer Array (SKA) project 177
  see also data science; machine learning

spatial statistical methods 240
Sri Lanka 146, 149, 153–156
statistical capacity 99
storage of data 192
structural data justice 254–256
  model of 256
    encompassing datafication 262–264
    encompassing the impact of data systems on structure 260–262
    encompassing the impact of pre-existing structure 258–260
    field study evidence 256–257
    foundation model 257–258
    human agency 266
    revised model 264–266
    tests to develop a more comprehensive model 257–264
  views of 254
survey data 215–216
  digitally enabled survey methods 215–216
  poverty indices 118
Sustainable Development Goals (SDGs) 2, 38–39, 71–73, 90–92, 99
  African countries' leveraging of big data in their SDG efforts 77–78, 85, 86–87
  African Data Consensus 80
  big data for SDG measuring and monitoring 92–93
  Global Partnership on Sustainable Development Data 10, 137
  human needs-based approach 77
  Kenya's open-data initiatives 78–79
  LIRNEasia 155
  methodological approaches to the study of data 236
  sources of big data for research C 208
  Tier Classification Criteria 93
  Uganda 158
  use of big data 76–77

Telefónica 272–273
Telenor 276–280
temporal detail, increased 42
theory of change 91, 102
three functions of data 105–107
time series analysis 237
traditional methodological approaches 226–228, 231–232, 235–237
traditional statistical methods 235–237
transport/ transportation planning 149, 154, 208, 236, 256–257, 261–262
Traunmueller, M. 64–65
"tyranny of algorithms" 139

Uganda
  accessing and sharing information via public radio discourse 160–161
  access to digital devices 159
  analyzing public radio content with AI technology 161–163, 168
    biases 168
    challenges 168–169
    methodology 162–165
    representativeness 168
    results 164–168
  Human Development Index 158
  opportunities and challenges of applying big data and AI to promote human development 159–160
  results of public radio content analysis 164, 168
    public sector accountability 167
    quality of public service delivery 164–165
    societal behavioral change 166
  Sustainable Development Goals 158
United Kingdom 14–15, 42, 59
United Nations (UN)
  Open Governance Partnership (OGP) 79–80
  Sustainable Development Goals *see* Sustainable Development Goals
  United Nations General Assembly (UNGA) 79
  United Nations Global Pulse 10–11, 53, 118–119, 159–160, 162, 273
  United Nations Economic Commission for Africa (UNECA) 71–72
  World Data Forum 12, 80–81
United Nations Development Group 235–236
United States 54–55, 141

World Bank 36–37, 117
  Statistical Capacity Index (SCI)
    correlation between SCI and HDI 99–101
Young, I. M. 254