CHAPTER 4

# Creative reuses of data for greater value

## Main messages

1. Innovations in repurposing and combining public intent and private intent data are opening doors to development impacts previously unimaginable. These innovations can inform and advance policy goals, help governments improve and target service delivery, and empower individuals and civil society.

2. When private intent data are repurposed for public purposes, they can help fill data gaps and provide real-time and finer-scale insights. When public intent and private intent data are combined, some or many of the limitations of each data type can be overcome.

3. Private intent data can be difficult to understand, monitor, and regulate. They may also miss the poorest or other marginalized populations and perpetuate discrimination and biases. Data protection is a key issue. Responsive regulation and consumer protection measures are needed, along with recognition of which populations are omitted from an analysis.

4. Using private intent data for effective policy making requires short- and long-term coordinated investments in training, data partnerships, and research. Best practices and guidelines need to be developed.

## The power of repurposing and combining different types and sources of data

Lack of data and information is no more apparent than during a crisis such as the COVID-19 pandemic or an earthquake. Urgent questions—What is happening? How can we help?—should receive good answers, and right away.

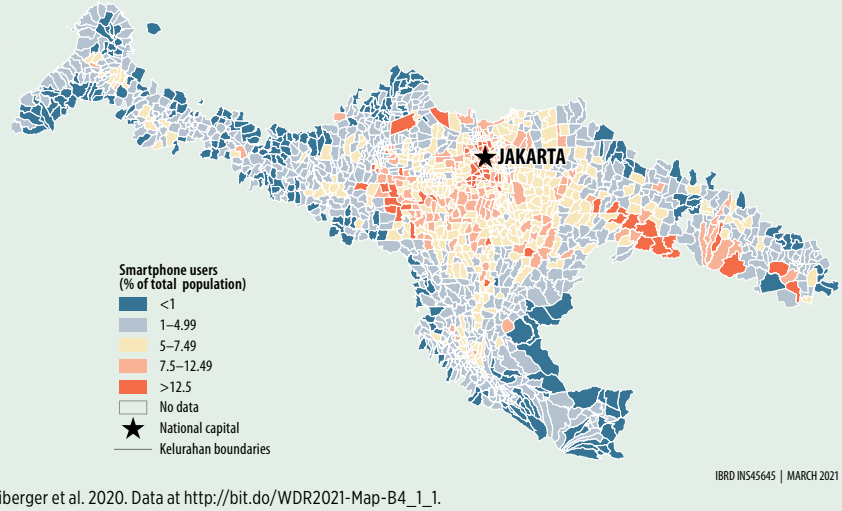
Consider the earthquake that devastated Haiti in 2010. Large donations of supplies and money poured into the country within days of the disaster, but delivering relief was difficult because vast numbers of people scattered. Censuses were no longer useful in helping responders direct relief to the people who needed it most. Using data from mobile phones, researchers were later able to demonstrate that they could have pinpointed population movements in almost real time. They found that one-third of the estimated 630,000 residents of the capital, Port-au-Prince, had fled the city.[1] Even though this study was retrospective, it demonstrated how real-time, spatially pinpointed information like this could have expedited relief efforts and saved countless lives had it been accessed contemporaneously. This example highlights an emerging question in development research: When a pressing crisis such as the Haiti earthquake or the COVID-19 pandemic emerges, what data can complement traditional public intent data to solve complex development challenges?

Recent technological shifts in lower-income countries—such as the adoption of mobile phones, social media, digital transactions, and mobile money—have generated a wealth of granular private intent data (see chapter 3 and box 4.1) suited to a wide range of secondary uses.[2] These data are being leveraged to

### Box 4.1 Using cellphones to combat COVID-19

After the onset of the COVID-19 outbreak, governments began implementing policy measures to reduce social contact and curb the spread of the pandemic. Data collected through mobile phones, such as call detail records and global positioning system (GPS) location data, have been extremely valuable in quantifying the effectiveness of policies, ranging from partial curfews to strict lockdowns. These data enable measurement of population density, travel patterns, and population mixing in real time and at high resolution, making it possible to better target policy interventions and improve epidemiological modeling.[a] Analysis of GPS locations showed that by March 23, 2020, social distancing policies had helped reduce mobility in major US cities by half.[b] In Colombia, Indonesia, and Mexico, the impact of nonpharmaceutical interventions (such as travel restrictions and lockdowns) on mobility differed by socioeconomic group. Smartphone users living in the top 20 percent wealthiest neighborhoods in Jakarta, Indonesia, reduced their mobility up to twice as much as those living in the bottom 40 percent

#### Map B4.1.1 Mapping the home location of smartphone users in Jakarta, 2020



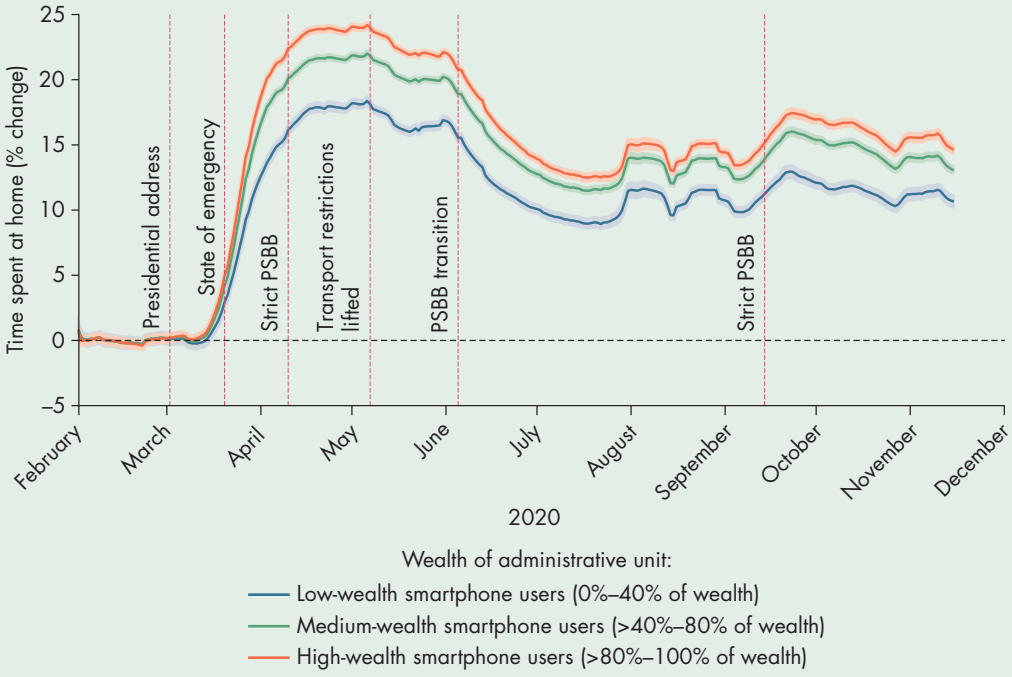

*Source:* Fraiberger et al. 2020. Data at http://bit.do/WDR2021-Map-B4_1_1.

*Note:* This map of Jakarta's metropolitan area shows the spatial distribution of smartphone users' home location as a percentage of Jakarta's total population.

*(Box continues next page)*

## Box 4.1 Using cellphones to combat COVID-19 *(continued)*

### Figure B4.1.1 Smartphone location data reveal the changes in the time users spend at home in Jakarta




*Source:* Adapted from Fraiberger et al. 2020. Data at http://bit.do/WDR2021-Fig-B4_1_1.

*Note:* Figure shows the changes in the time users spent at home from February 1 to November 15, 2020, relative to the baseline period. PSBB = *Pembatasan Sosial Berskala Besar* (large-scale social restrictions).

(map B4.1.1 and figure B4.1.1).[c] Using an epidemiological model and estimates of population movements derived from mobile phone data, research in China found that nonpharmaceutical interventions implemented in late January 2020 led to a 98.5 percent reduction in the number of COVID-19 cases one month later.

Meanwhile, mobile phones have proved to be a valuable tool for contact tracers seeking to alert individuals who may have been in contact with an infected person.[d]

Although both private companies and government actors have produced mobile phone applications for contact tracing (such as the Corona app 100m in the Republic of Korea, TraceTogether in Singapore, and COVIDSafe in Australia), their efficacy relative to more traditional forms of contact tracing has not yet been established. Digital contact tracing also raises important concerns about data protection,[e] prompting researchers worldwide to develop contact tracing technologies that preserve privacy. Examples are the Private Kit: Safe Paths developed by the Massachusetts Institute of Technology (MIT) and the Decentralized Privacy-Preserving Proximity Tracing (DP3T) protocol developed by a consortium of European research institutions.

Despite the potential of deploying mobility data in the fight against COVID-19, their impact on policy thus far has been limited, especially in lower-income countries. Bottlenecks include a lack of technical expertise among government organizations; restrictions on data access, especially by mobile network operators; and lack of the investments and political will required to scale up one-time projects.[f] To ensure that mobility data can be made accessible and useful for policy purposes, it is important for all stakeholders—governments, mobile phone operators, technology companies, and researchers—to collaborate and form interdisciplinary teams to facilitate readiness and responsiveness to future crises.

a. Buckee et al. (2020).
b. Klein et al. (2020).
c. Fraiberger et al. (2020).
d. Servick (2020b).
e. Servick (2020a).
f. Oliver et al. (2020).

monitor the effectiveness of policy measures and predict outcomes of long-standing concern to development practitioners and policy makers.

Technological advances in the private sector have turned data into an integral component of the production process, leading to gains in productivity and generating even more data that can be repurposed for development. Specifically, the same approaches that are transforming efficiency and innovation in the private realm are being repurposed to tackle development bottlenecks in poor countries, making the development process more efficient, innovative, agile, and flexible. Because of the nonrivalrous nature of data, private companies also are able to reuse and repurpose publicly collected data, which can generate welfare-enhancing economies of scope.[3]

That said, the reuse of private intent data is not a panacea and may pose unique challenges for policy making. For example, data created by businesses to track mobile phone users may miss the poorest populations who do not have these technologies. Similarly, the data required to target customer experiences and to achieve business gains are different from the sociodemographic information on which policy makers rely to design inclusive policy. Furthermore, many of the algorithms used to process private intent data are considered trade secrets and thus lack the transparency required for effective policy making. Transparency and oversight are also important considerations when giving private companies access to sensitive data such as those related to facial recognition and surveillance (see chapter 6).

Despite these challenges, combining public intent and private intent data can offer real-time insights that not only are inclusive of the entire population (or nearly so) but also are more precisely estimated for specific population segments and localities. This is especially important for the poorest people in the poorest countries, which have the largest data gaps. Too often, individuals on the lowest end of the income distribution remain on the margins when government, civil society, and the private sector lack the data to effectively allocate and target resources based on need. Leveraging all available data may reveal insights for the poor and marginalized that were previously unattainable.

This chapter begins by showcasing innovative uses of public intent and private intent data for aiding development policy. Examples include data repurposing and synergies to improve predictions of disease spread, streamline service delivery, and allocate aid in disaster recovery. The chapter then turns to an exploration of the challenges that arise when private intent data are repurposed or when public intent and private intent data are combined. It concludes with a framework within which policy makers and funders could invest in the human capital, data partnerships, and research needed to gain useful insights from these new types and combinations of data.

## Features of private intent data that can overcome gaps in public intent data

Private intent data are an alluring candidate to overcome public intent data gaps and offer new perspectives on development problems. These types of data are increasingly large in scale, "always on," zoomed in, and, at times, less biased.

*Big data.* Private intent data are typically labeled "big data,"[4] recognizing their wide reach and scope. The growing rates of mobile phone and social media usage enable information to be gathered from all users on these platforms. Although this process may underrepresent certain parts of the population in countries with lower usage rates, ever-larger portions of a population are being brought into the fold as the rates of mobile phone ownership and internet connectivity continue to increase, even in lower-income countries. When private intent data are repurposed toward a public goal, their volume and reach can not only inform first-order policy goals of poverty reduction and service delivery, but also facilitate efforts to detect and study rare events, such as fraud, corruption, or criminal activity, through techniques such as anomaly detection.

*"Always on" data.* Private intent data are always on[5] because the daily use of new technologies entails constant data collection. Call detail records (CDRs) and apps that log locations pinpointed by satellite-based global positioning systems (GPS) offer traces of where cellphone users travel throughout the day. When a sudden and unexpected shock hits, such as a natural disaster or a disease outbreak, such data can provide precious real-time information on human mobility and call density. The timeliness of private intent data therefore contrasts with public intent data, which are generally collected at intervals of 1, 5, or 10 years and thus are not always very timely. In Africa, for example, 14 of 59 countries did not conduct any surveys from 2000 to 2010, impeding the construction of nationally representative poverty measures.[6] This critical situation sparked the call for a "data revolution" by the United Nations in 2014, pushing for an increase in data collection efforts in Africa and elsewhere.[7] Although the situation is improving, with the average number of surveys per country per year increasing from 0.5 in 1990 to 1.5 in 2010,[8] the

lack of timeliness of public intent data has resulted in huge knowledge gaps, which are particularly glaring following major economic shocks such as COVID-19. Meanwhile, private intent data are increasingly being used to help fill these gaps.

*"Zoomed in" data.* Private intent data can zoom in on individuals and locations. Private companies want to know who is using their products or services and in what ways they can optimize their offerings and operations. Private intent data zoom into individuals to collect key metrics such as transaction histories to predict consumer behavior and bolster successful products. Internet Protocol (IP) addresses, browsing histories, and smartphone app logs add to a rich dataset that companies collect on a single person over time. Tracking whether app users enter a store or whether IP addresses in a neighborhood are searching for products on their site enable companies to better plan their store locations and stock their supplies. These data are now being applied to the public sphere, ranging from improving population maps[9] to helping decision-makers target and optimize critical development resources. A key challenge to using individual data patterns to allocate resources or establish eligibility for products and services is data manipulation: individuals may strategically change browsing or other data usage behavior to appear more favorable in ranking criteria used by data algorithms to make allocation decisions. More research and policy deliberations are needed to design algorithms and decision rules that account for such user manipulation.

*Potentially less biased data.* Private intent data potentially reveal less "biased" information about people than surveys or polls because researchers observe actual behavior instead of relying on responses. Although it is possible that respondents misreport answers during surveys, they have little incentive to do so when searching the internet. For that reason, the Google internet search engine has been dubbed a "digital truth serum."[10] This finding may apply especially to opinions on sensitive topics such as racism. Few will admit their opinions in surveys, but they are revealed through internet searches and can influence political outcomes, among others.[11] However, the algorithms used by search engines are considered private trade secrets and are usually optimized for private benefit—not public benefit. Without knowledge of the workings of these algorithms, users of search engine outcomes as an exclusive source of data may find they lead to biased and discriminatory policy predictions.

Overall, combining public intent and private intent data is a powerful way to gain aggregate population insights in real time, if enough attention is given to addressing representativeness, discrimination, and transparency. Calibrating private intent data with census and survey data is one way to estimate population-level needs.

The next section offers a broad range of innovative examples of applications of private intent data to public policy and instances in which public intent and private intent data have been combined to promote inclusive and timely development solutions.

## New insights from repurposing and combining data

The last decade has seen a surge in innovative research that repurposes private intent data and combines it with public intent data to tackle development issues. In the spring and summer of 2020 when the COVID-19 outbreak reached global dimensions, more than 950 scientific and medical articles were published that used private intent data to tackle the pandemic (box 4.2). Researchers' ability to respond quickly to the pandemic builds on a growing trend of research that combines diverse data to tackle emerging issues.

### Monitoring public health

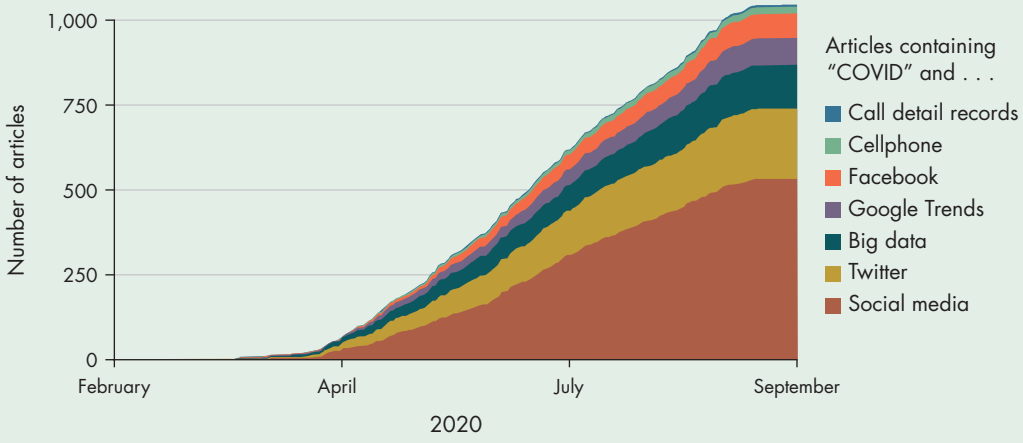
Monitoring public health is a key area that could benefit from repurposing and combining public intent and private intent data. In many lower-income countries, infectious diseases routinely pose large health threats. Five of the top 10 causes of death in low-income countries are communicable diseases, including lower respiratory infections, diarrheal diseases, HIV/AIDS, malaria, and tuberculosis.[12] Viruses have been responsible for more deaths than all armed conflicts around the world over the last century.[13] Especially in countries where data are limited, new big private intent data sources can help inform public policy interventions to reduce the mortality and morbidity rates from infectious diseases. Identification of hotspots can help disease control programs target activities more effectively to those areas, reducing infection rates both directly and indirectly in destination areas that are receiving infected travelers.[14]

As early as 2008, researchers began exploring how mobile phone data could be used to measure population mobility and then be applied to the study of epidemics.[15] A seminal study applied this research at scale for all of Kenya using mobile phone data on nearly 15 million individuals to identify sources of imported malaria infections stemming from human mobility.[16] During the 2014 Ebola outbreak in West Africa, researchers highlighted the potential benefits

## Box 4.2 Leveraging private intent data to tackle COVID-19

Between February and September 2020, more than 950 articles were published in scientific, medical, and technical journals that repurposed cellphone, social media, Google search, and other types of big private intent data to better understand the spread of COVID-19 and to offer policy and operational solutions (figure B4.2.1). Despite the relatively large number of articles in a short time span, coverage of lower-income countries was low, especially those in Africa (map B4.2.1). Lack of expertise, poor training, difficult access to data, and limited research support are key areas that funders could address to ensure innovative uses of data in and about lower-income countries.

### Figure B4.2.1 Use of repurposed data to study COVID-19: Published articles, by type of private intent data used




*Source:* WDR 2021 team, based on data from CORD-19 (COVID-19 Open Research Dataset) Semantic Scholar team, Ai2 (Allen Institute for AI), http://www.semanticscholar.org/cord19. Data at http://bit.do/WDR2021-Fig-B4_2_1.

*Note:* Figure shows the number of articles published in scientific, medical, and technical journals across time from February to September 2020. The cumulative sum across all categories is higher because some articles appear in more than one category.

*(Box continues next page)*

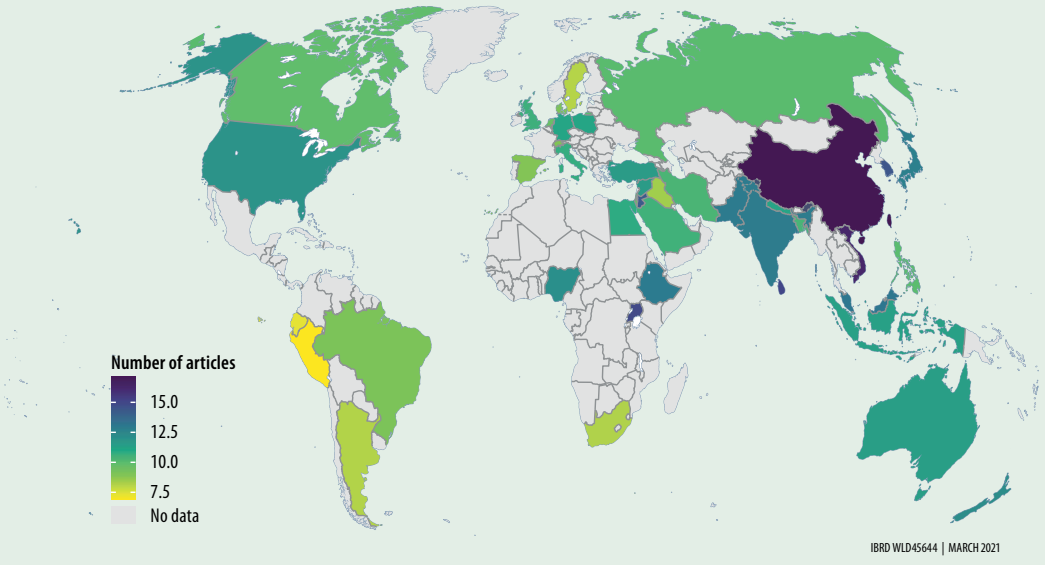
of using mobile phone data in the design of public policy.[17] However, use of these analytics at the time of the crisis remained limited.[18]

After onset of the COVID-19 pandemic, countries began to deploy this type of research and to pair mobile phone data with public intent data. Belgium formed a Data Against COVID-19 task force to analyze deidentified mobile phone data. These data are being used to monitor changes in human mobility trends due to lockdown measures and to inform decisions related to appropriate lockdown measures. In the Republic of Korea, mobile phone data are being used to aid contact tracing efforts to contain disease spread. By combining mobile phone data with medical facility records, credit card transaction logs, and closed-circuit television recordings, the government is identifying people at risk of exposure.[19] Lower-income countries such as Ghana and Mozambique are beginning to use deidentified mobile phone data to combat the pandemic, typically with the support of international organizations that provide analytical skills for processing the data.[20]

Other types of big data are also being enlisted to create measures of mobility that can improve the effectiveness of the pandemic response. Facebook disease prevention maps are being used to study COVID-19 and have been expanded to include colocation maps that measure comingling among people living in different areas and trends in whether individuals are staying near their homes or continuing to go to other locations.[21] Google has produced a new set of measures to track the response to policies aimed at flattening the curve of the COVID-19 pandemic.[22] Other sources of data for GPS locations have been

## Box 4.2 Leveraging private intent data to tackle COVID-19 *(continued)*

### Map B4.2.1 Uses of repurposed data to study COVID-19: Published articles, by country




*Source:* WDR 2021 team, based on data from CORD-19 (COVID-19 Open Research Dataset) Semantic Scholar team, Ai2 (Allen Institute for AI), http://www.semanticscholar.org/cord19. Data at http://bit.do/WDR2021-Map-B4_2_1.

*Note:* Map shows the number of articles published in scientific, medical, and technical journals across countries from February to September 2020. Article counts are divided by the COVID-19 death incidence rate.

used by data analytics firms such as Baidu, Cuebiq, and Unacast to assess the impacts of social distancing measures for COVID-19.[23] GPS data provide better approximation of locations and mobility at a finer spatial resolution, but their availability is limited by smartphone penetration and usage. In many lower-income countries, smartphone penetration is still low, and even those individuals with smartphones may only selectively turn on data or GPS because of high costs and drain on battery life.

The potential of new data sources for supporting public health and epidemiology efforts goes far beyond measures of mobility.[24] Efforts are under way to use data tools as early warning systems for outbreaks and for understanding disease dynamics and routes of transmission. For example, the company BlueDot provides infectious disease surveillance services using advanced data analytics. It was able to warn of the outbreak of COVID-19 before the official announcement in early January 2020 by analyzing news reports, disease networks, and official proclamations.[25] A similar prediction was made for the 2015–16 Zika outbreak that affected an estimated 1 million people, mainly in Latin America.[26] By combining online news sources, Google search queries, Twitter posts, and government disease reports, local outbreaks could have been detected two to three weeks earlier, a retrospective study estimates.[27] Combining public intent and private intent data sources has also improved forecasts for Ebola in West Africa[28] and dengue in Southeast Asia.[29] Improved forecasting of disease outbreaks and associated population movements is essential for efficient response measures to curb incidence rates.[30]

Another open and fertile source of synergy is data collected by wearables and other biotech devices. For example, the Kinsa HealthWeather app tracks fevers around the United States via smart thermometers and uses the aggregate data to create prediction models for the spread of disease. This type of application is particularly relevant in crises such as COVID-19, where timely reporting of case growth can help

accurately map disease spread and enable timely and appropriate public policy responses.

### Targeting resource allocations and responses during crises

Approximately 20–30 million people worldwide are displaced every year because of natural disasters such as storms, floods, droughts, and geological events.[31] Over the last decade, about 600,000 people lost their lives to natural disasters, most of them in low- and middle-income countries.[32] Effective disaster prevention, mitigation, response, and recovery require timely, cost-effective data at fine spatial scales. However, many countries lack the adequate early warning systems and advanced geological tools to aid in this process—at times with devastating consequences. During the 2018 earthquake and tsunami in Central Sulawesi, Indonesia, the government could have minimized the human cost had the country's warning system of buoys and seismographic sensors not been defective.[33] As climate change continues to increase the frequency and damage of natural disasters, lower-income countries will likely bear the brunt of the economic and human impacts. Spotlight 4.1 highlights the importance of improved meteorological data for lower-income countries to confront enhanced climate risks.

Recent data innovations have revealed that nontraditional sources of private intent data such as mobile phone usage, social media activity, online queries, crowdsourcing platforms, and remote sensing technologies can facilitate disaster management.[34] These devices and activities are not a replacement for advanced geological and meteorological equipment, which can predict disasters and offer early warnings. They can, however, help in government efforts to prevent loss and provide relief when such events occur. Various studies in both lower- and higher-income countries have found that scraping social media platforms for posts related to seismic activity produces an in situ impact profile of seismic damage similar to the ones produced by advanced geological instruments, the traditional source of such data.[35] Similarly, Tweets have been analyzed for disaster-related keywords to detect earthquakes in Australia and New Zealand.[36] Deidentified CDR data are a good predictor of population movement for weather-related disasters such as floods. For example, the textual content of Tweets was used to understand how people were reacting to the 2011 floods in Thailand. Messages were classified by their content to help highlight precise needs in affected communities.[37]

The geospatial nature of social media posts can further help prioritize resource allocation in times of dire need. Moreover, combining geographic and social media analytics can enhance aid recovery efforts after a disaster. In the aftermath of the 2014 earthquake in Napa, California, researchers trained a machine learning algorithm to extract disaster-related semantics from Tweets and paired this information with geolocations to identify spatial hotspots.[38] From these data, they were able to infer a disaster footprint and assess damage. They also learned that this method was transferrable to other social media platforms and locations, with tweaks for cultural differences in social media use. Similarly, researchers studying Hurricane Irma, which hit Florida in 2017, found that sentiment analysis[39] on geolocated Tweets could be used to guide resource allocation.[40] Social media and mobile records have also proven useful in tracking recovery efforts. After Hurricane Sandy slammed into the New York City area in 2012, researchers analyzed Tweet topics and sentiment to see how those who experienced the disaster were coping, compared with those who did not experience it.

Finally, governments have long used satellite imagery to assess damage in the aftermath of natural disasters. However, this imagery usually lacks the spatial resolution needed for a granular assessment. It is typically considered public intent data, but a growing number of private companies are launching their own remote sensing technologies and data collection. The start-up Cloud to Street uses private satellite data to provide near real-time flood assessments to assist disaster recovery and adaptive planning. In three days in 2018, it was able to build a flood monitoring system to help the Democratic Republic of Congo deploy resources to 16,000 asylum seekers who had sought refuge along the flood-prone banks of the Congo River. Cloud to Street leveraged high-resolution private intent satellite data with data about cropland, population, and public assets (such as roads and infrastructure) to generate real-time impact estimates served on an interactive web platform and with automated alerts. As decision-makers transitioned from disaster response to recovery, Cloud to Street transitioned to using freely available satellite images—an effort that enabled longer-term support with fewer resources.[41]

### Mapping poverty and targeting service delivery more precisely

Timely, reliable data on population characteristics are vital for responsive social and economic policy making. Mobile CDR and remote sensing data have

recently been used to predict poverty patterns on a granular level and in a timely fashion, thereby helping to better target government services. Use of these data sources costs a fraction of that for fielding censuses or household surveys. Similar data from social media, online engagement, and satellite imagery are reducing the constraints to collecting data on the most vulnerable and hard-to-reach populations. Moreover, the same algorithms that Google and Facebook use for online consumer marketing can be tweaked to direct resources to people living in poverty. In the same way that these tech firms predict the advertising that may interest consumers based on their digital behavior, development actors can use digital behavior to predict whether people are economically vulnerable.[42]

Research relying on data from Rwanda reveals that past histories of mobile phone use extracted from CDRs are a reliable predictor of socioeconomic status as validated against survey data.[43] Moreover, the researchers find that the predicted characteristics of millions of mobile phone users can be aggregated to the same distribution of wealth across the entire country or at the cluster level—approximately equivalent to a village in rural areas or a ward in urban areas—as that indicated by traditional data sources. Such highly localized poverty maps can be used to effectively target policies, programs, and resources to the poorest. These methods can also improve demographic targeting of services by gender, age, and income level. For example, CDR data have been used to identify the gender of phone users,[44] as well as to identify the ultrapoor.[45]

Beyond the realm of CDRs, research in higher-income countries has shown that online browsing history and social media activity can also reliably predict household income. Social media footprints were used in Spain to infer city-level behavioral measures and predict socioeconomic output, specifically unemployment.[46] Similarly, data from Yelp reviews of retail shops were used to measure changes in gentrification and predict local housing prices.[47] Equipped with real-time and localized insights and trends, policy makers can better inform policies to target areas that have been affected by short-term economic shocks or long-term economic shifts.

Remote sensing technology is yet another novel way to collect population characteristics, predict poverty patterns, and improve public service delivery.[48] Researchers have relied on publicly available data from Africa to both calibrate and validate machine learning models. The Demographic and Health Survey (DHS) sponsored by the United States Agency for International Development (USAID) and the World Bank's Living Standards Measurement Study (LSMS) surveys provide high-resolution data on household wealth and consumption expenditures. When calibrated with these surveys, satellite imagery can predict poverty. At the survey cluster level, when used with survey data from Malawi, Nigeria, Rwanda, Tanzania, and Uganda satellite imagery can explain 55–75 percent of the variation in wealth and consumption per capita. Estimates of economic well-being using this approach outperformed both similar estimates using satellite readings of nighttime light in the same countries and estimates using mobile phone data in Rwanda. Critically, this approach has been shown to work reasonably well for predicting wealth and poverty in countries when they are excluded from the sample used to train the model, suggesting the approach is scalable across other countries, at least in Africa.

### Ensuring road safety in transport and transit

Road transport is an important element of economic development. Access to transport and mobility are highly correlated with income and quality of life. Even though lower-income countries have only half of the world's vehicles, they account for 90 percent of road traffic fatalities. In 2011 the World Health Organization (WHO) and the World Bank launched a Decade of Action for Road Safety, and they have provided funding and technical assistance to build systems aimed at reducing injuries and deaths on the road. Despite these efforts, little progress has been reported in low- and middle-income countries, and the number of fatalities remains high.[49]

A new and growing body of literature studies how alternative sources of data can be used to make progress toward achieving national road safety outcomes. In the public sector, for example, a study in Nigeria provided road safety agents with a monitoring system to investigate and record road safety events via mobile phone.[50] Access to this mobile phone–based database helped disseminate information better and enabled agents to respond faster to road accidents. Such transit monitoring practices are becoming more widespread, especially in the private sector. Commercial banks in Kenya now require a tracking device in minibuses before approving loans to bus service owners. As a result, today most long-range buses in the country are equipped with GPS.[51] This technology advancement serves the dual purpose of tracking assets under lien for the bank's private benefit and promoting safer driving for public benefit.

Social media analytics have also been applied in the private sector to understand the traffic safety culture. A recent study in Washington State in the United States mined Twitter data to understand the patterns, behaviors, and attitudes related to road safety.[52] The study conducted sentiment analysis based on traffic-related keywords to extract latent views on topics such as safe driving measures, accidents, law enforcement and patrolling, and accident-causing behavior. It found that sentiment analysis using social media posts can be used in developing policies to improve traffic safety relevant to specific contexts. This type of sentiment analysis could be applied in lower-income countries as well, with substantial benefits. Techniques are also being developed to fill in gaps in data on the number and location of accidents in lower-income countries. Recently, researchers developed an algorithm to identify and geolocate crashes from Twitter feeds to substantially increase the digital data available to prioritize road safety policies. Spotlight 4.2 describes how car crash danger zones were pinpointed in Nairobi, Kenya, by combining police reports and crowdsourced data.[53]

More broadly, research in this area has focused on the transit industry to answer broader development questions in the realm of private sector development. For example, a study in Kenya found that providing bus owners with data on their employees' driving behavior can improve firm operations.[54] Specifically, they placed GPS devices in Kenya's inner-city public transport vehicles and tracked a variety of data that captured driving behavior, including acceleration, jerk, location, and timestamp to measure the number of daily safety violations. The main contribution of this data innovation was to correct informational asymmetry: once minibus owners could track driving performance, drivers could receive more generous contracts for better performance. In turn, drivers operated in a manner less damaging to the vehicle, more frequently met targets, and reduced underreporting of revenues. Thus incentives between the company (principal) and the drivers (agents) were better aligned. These types of data can also provide governments with feedback to use in redesigning their road infrastructure and guide interventions to reduce accidents.

### Monitoring illegal fishing and deforestation

Recent advances in combining public intent and private intent data are also improving the monitoring of natural resource extraction. Box 4.3 features one example: identifying illegal fishing in protected ocean waters.

Efforts to monitor deforestation have also begun to leverage public and private datasets. Combining data in this way has enabled indigenous groups to patrol their forest reserves and defend against encroachment. With the aid of open-access or cheaper private satellite imagery, cloud computing, community observations, and publicly available property maps, community-based forest monitoring has become increasingly effective in identifying encroachment.[55] In addition, through social media and platforms such as Global Forest Watch the international community can better help local groups hold governments accountable in achieving national sustainable development commitments.[56] Similar data are being used by companies to ensure that their suppliers are meeting sustainability standards for forest products. A recent initiative, Radar Alerts for Detecting Deforestation (RADD), was launched by the world's 10 largest palm oil producers and buyers to monitor illegal deforestation in palm oil plantations.[57] By funding development of a system to detect illegal deforestation using public radar imagery, property maps, and private procurement data, this initiative may signal a shift from civil society monitoring the private sector to the private sector monitoring itself to ensure that company commitments are met.

### Keeping governments accountable

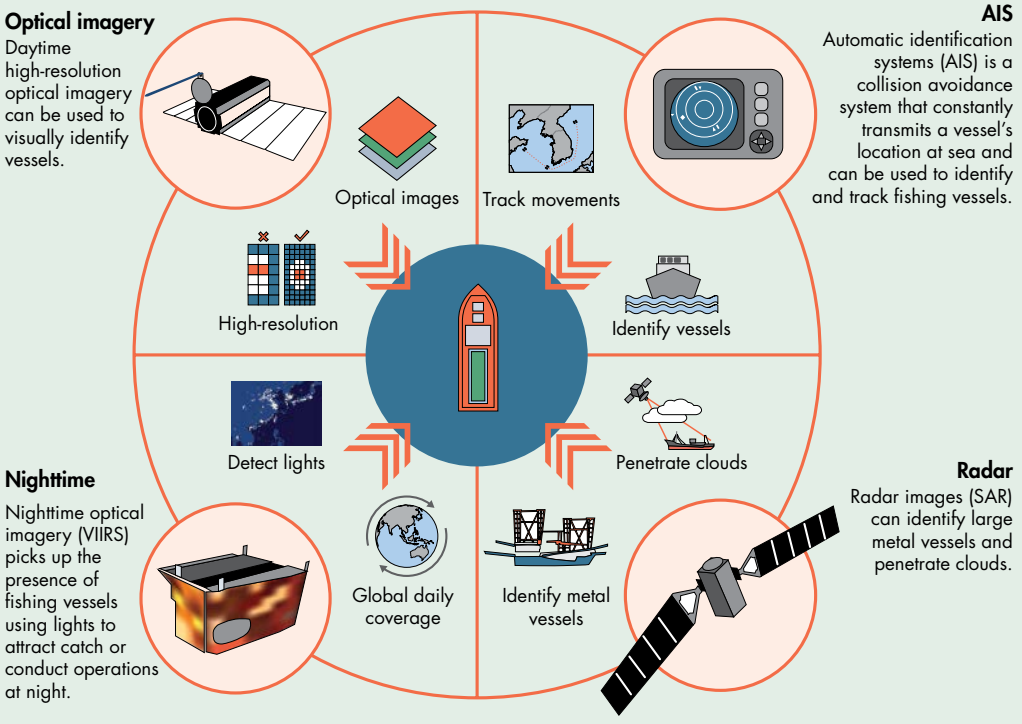
Emerging data types are enabling civil society to better monitor corruption. Utilizing crowdsourced data and web scraping, social media discussion boards are emerging as ways in which local leaders can act against corrupt officials and receive real-time feedback on the impact of anticorruption policies.[58] Data reported in newspapers have been used to target corruption, thereby allowing civil society organizations to press for stricter governance measures. A systematic, real-time view of corruption trends can be gained from the news flow indices of corruption (NIC) constructed by the International Monetary Fund (IMF), drawing on country-specific searches of more than 665 million news articles.[59] Regressing the NIC onto the real per capita gross domestic product (GDP) revealed that changes in corruption levels as measured by the NIC indicators were associated with 3 percent lower economic growth over the next two years. Combined with election data, NIC data have helped identify countries that had peaks in corruption before or after elections. These findings can prove helpful to international responses to corruption.

Private sector data are making it possible for international organizations and civil society actors to monitor policy and report on important events

## Box 4.3 Preventing illegal fishing in protected maritime areas

Monitoring illegal fishing in Marine Protected Areas (MPAs) is difficult because of their size and distance from land. The boundaries of MPAs are curated and made open access by the United Nations Environment Programme (UNEP) and the International Union for Conservation of Nature (IUCN). Yet identifying boats in vast expanses of the ocean requires innovative uses of data that are not publicly available. Global Fishing Watch has data partnerships with the firm ORBCOMM to access raw data from commercial trawlers' automatic identification systems (AIS), which provides the real-time geographic coordinates of each trawler to help avoid collisions and provide other traffic services. AIS data can be combined with optical and radar imagery from satellites to detect illegal fishing activity (figure B4.3.1). By overlaying MPA boundaries on AIS data used to identify boats and determining fishing behaviors from the time spent in specific areas, researchers found that 59 percent of MPAs in the European Union were commercially trawled. In areas that were heavily fished, the presence of sensitive species (such as sharks, rays, and skates) was 69 percent lower.[a]

### Figure B4.3.1 Public intent and private intent data can be combined to detect illegal fishing activity




*Source:* Infographic taken on July 8, 2020, globalfishingwatch.org. © Global Fishing Watch. Used with permission of Global Fishing Watch; further permission required for reuse.

*Note:* Public intent data include satellite data. Private intent data include data from trawlers' collision avoidance systems.

a. Dureuil et al. (2018).

such as elections in real time. The Inter-American Development Bank, in partnership with governments in Latin America, has launched a website that uses crowdsourced civic feedback to monitor public works projects.[60] Similarly, Civic Cops, a start-up in India, provides a suite of digital platforms to connect governments with civil society, notably offering a service that allows civic complaints and citizen service requests to be filed by mobile phone and directed to the corresponding public authorities. Civic engagement data

have also been used to monitor elections in lower-income countries. For example, in Sierra Leone's 2012 elections a collection of citizen journalists traveled throughout the country and reported election activity through SMS text messages, which were then posted on a Tumblr website, pegged to a Google map, and disseminated on Twitter.[61]

### Benchmarking policy priorities

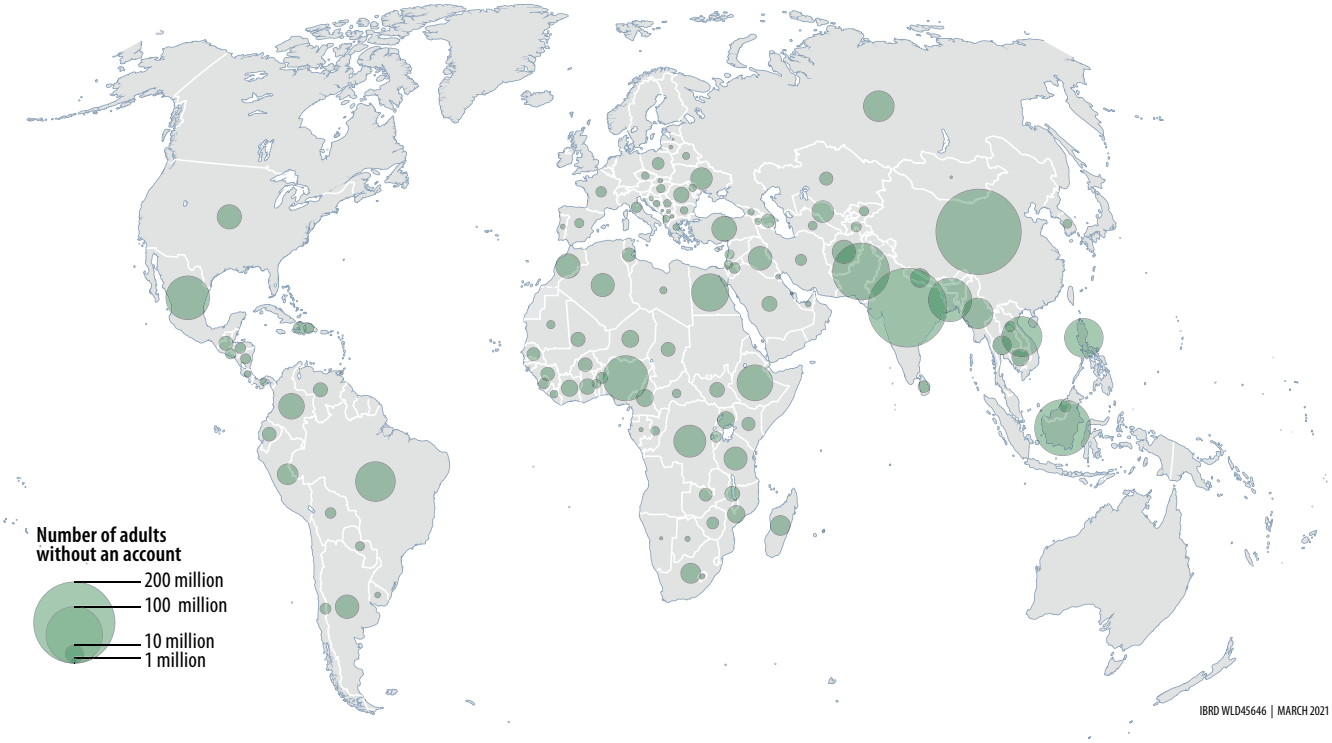
Private intent data repurposed by international organizations, civil society actors, and private companies are being used to track policy goals and benchmark policy priorities. These initiatives are invaluable because they provide unique and comparable data across countries that are not collected by national governments.

The data being harvested and disseminated to promote financial inclusion have been widely recognized by policy makers as critical to reducing poverty and achieving inclusive economic growth. Partnering with the polling firm Gallup Inc., the World Bank launched the Global Findex database in 2011, the world's most comprehensive database on how adults save, borrow, make payments, and manage risks (map 4.1). This dataset was created by adding a module to the Gallup World Poll, which offers a standing global survey that produces comparable data across countries and across time. Researchers, private companies, and international organizations use these data to understand the lives of people everywhere.[62] The Global Findex database has become a mainstay of global efforts to promote financial inclusion. In addition to being widely cited by scholars and development practitioners, Global Findex data have been used to track progress toward the World Bank's goal of universal financial access by 2020 and the United Nations' Sustainable Development Goals (SDG Target 8.10).

Data synergies can also help in critical policy areas such as food security in both times of normality and crises such as the COVID-19 pandemic. The potential

**Map 4.1 Private intent data can provide unique and comparable information not collected by national governments, such as the number of adults who lack a formal financial account**

*Globally, 1.7 billion adults lacked a formal financial account in 2017*




*Source:* World Bank, Global Findex (Global Financial Inclusion Database), https://globalfindex.worldbank.org/. Data at http://bit.do/WDR2021-Map-4_1.

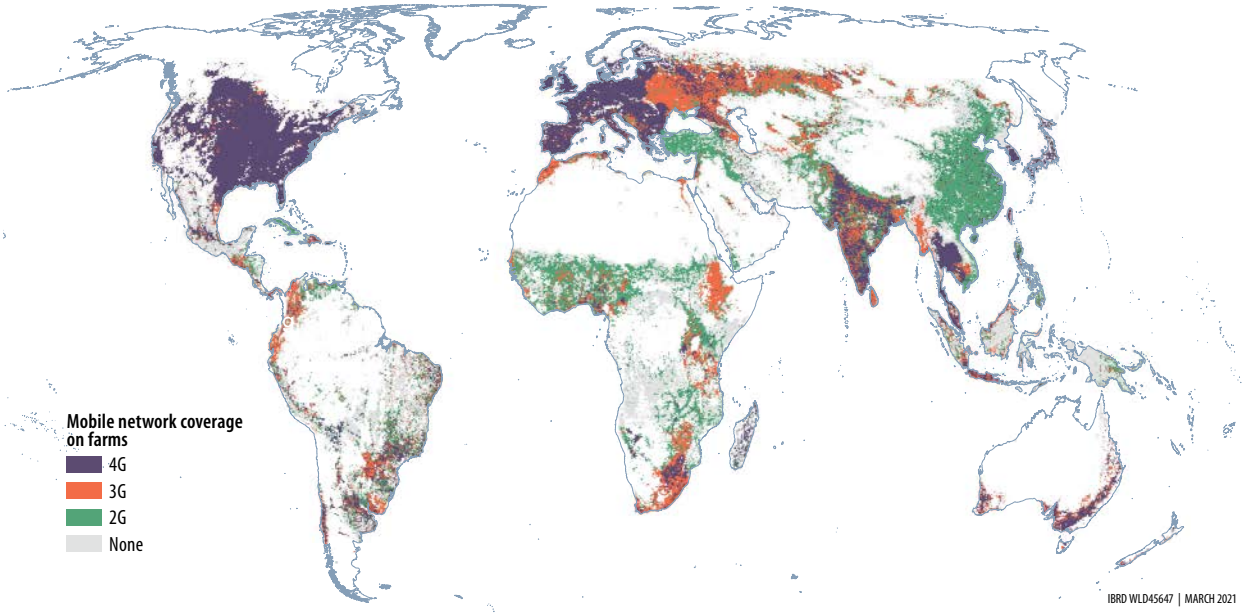
*Note:* Data are not displayed for economies in which the share of adults without an account is 5 percent or less.

to combine geospatial data with farmer output and market pricing can improve the logistics and management of critical food systems. Meanwhile, international organizations have partnered with companies to create public intent surveys to track progress toward the SDGs and inclusive development. For example, in 2014 the United Nations Food and Agriculture Organization (FAO) began to add questions to the Gallup World Poll to collect data for its Food Insecurity Experience Scale (addressing SDG 2). In 2015 the International Labour Organization (ILO) and Walk Free Foundation added questions that measure the incidence of modern slavery (addressing SDG Target 8.7). Through a partnership with Facebook, in 2018 the World Bank and the Organisation for Economic Co-Operation and Development (OECD) launched the Future of Business biannual survey.[63] The survey targets active micro, small, and medium enterprises (MSMEs) that host a Facebook business page. Using these data, researchers have been able to study the gender pay gap across 97 countries.[64]

Apart from surveys, companies are beginning to repurpose their own data for the public good. During the COVID-19 pandemic, Google began releasing updated community mobility reports for 135 countries.[65] These reports rely on users' location data to show daily changes in mobility patterns at the country or state/provincial level, such as fewer trips to transit stations, retail stores, parks, grocery stores, pharmacies, workplaces, or residential addresses. These data give public health officials and the general public a way to benchmark a region's response to COVID-19 relative to other regions and over time. Because the data are collected systematically across countries, they can also be used to compare behavioral responses across the world. Another example of a private company repurposing its own data for public benefit is the internet speed test company Ookla, which provides a global index for internet speeds that ranks countries for their mobile and fixed broadband.[66] These data can be used by governments and funders to prioritize investments in broadband coverage.

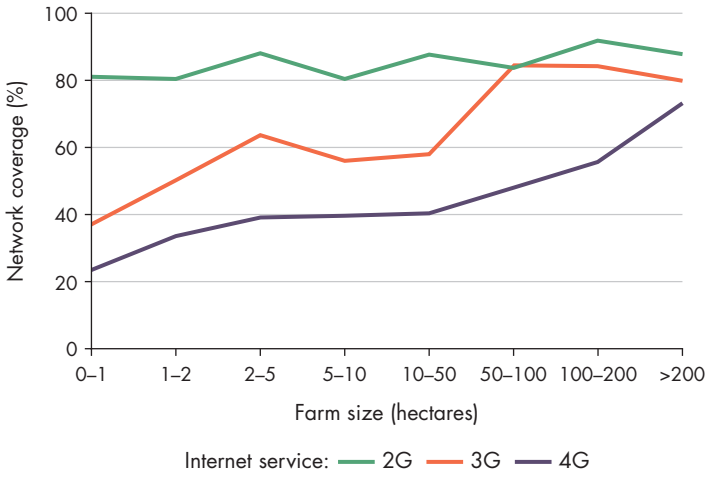
Researchers are also combining global public intent and private intent datasets to prioritize funding streams for donors. One example is in the digital agricultural space, where farmers can access extension services on their cellphones. Digital agricultural interventions offer a solution to the dearth of agricultural extension agents in many lower-income countries, where the ratio of farmers to extension agents often exceeds 1,000 to 1.[67] Digital services can provide farmers with expert scientific advice based on their local field, market, and climatic conditions. Yet most small-scale farmers live in areas with lower 3G and 4G coverage than in areas with relatively high shares of large-scale farms (map 4.2 and figure 4.1).[68] This

**Map 4.2** Agricultural extension services can be tailored to the slower, older broadband internet accessible to many small-scale farmers




*Source:* Mehrabi et al. 2020. Data at http://bit.do/WDR2021-Map-4_2.

**Figure 4.1** Gaps in network coverage differ across farm sizes, affecting agricultural extension services




*Source:* Mehrabi et al. 2020. Data at http://bit.do/WDR2021-Fig-4_1.

finding suggests that the wave of digital agricultural services should focus on 2G solutions (such as voice and text messaging) to ensure that small-scale farmers are reached. Combining private intent broadband coverage data from the data aggregation company Mosaik (now part of Ookla) with public intent farm size data yields localized estimates of broadband usage at 10 square kilometer resolution. This type of analysis can be used in making decisions about the deployment of infrastructure to support the faster broadband required for digital services that depend on smartphones.

## Limitations in using private intent data for development

Despite the enormous potential offered by private intent data through repurposing and synergies, several important limitations and challenges affect their use for development projects. These issues should be taken into account in the design of future research and public policy.

### Data coverage and representativeness

A key limitation of most private intent data is their lack of representativeness. Private intent data are often a by-product of the use of digital technologies such as mobile phones or the internet. Having access to these technologies typically requires infrastructure resources such as electricity or broadband that are distributed unequally in lower-income countries. In addition, because smartphone ownership is skewed toward those who can afford the phones, the data collected through these technologies primarily highlight the characteristics of a relatively wealthier share of the population. A 2012 study combining CDRs and surveys found that mobile phone owners in Rwanda were wealthier, better educated, and predominantly male.[69] Similar conclusions emerged from an analysis of the population of mobile phone owners in Kenya.[70] The lack of representativeness is even more pronounced in social media data, which typically require that users be literate in addition to having internet access. Moreover, because of the access charges associated with internet use, only the wealthy can afford to use the internet on their mobile devices. Estimates from Ghana, Kenya, Nigeria, and Senegal suggest that less than one-third of the population uses internet on a mobile phone, and less than 15 percent in Mozambique, Rwanda, Tanzania, and Uganda.[71] To overcome the lack of representativeness of private intent data, development practitioners often rely on statistical methods to combine them with public intent data.

One important source of alternative data is satellite imagery, which can be either public intent or private intent, depending on the application. Images collected by satellites have the advantage of being fully representative of the population, and they are well suited to picking up measures of building density that are highly correlated with population density and, by extension, economic well-being. Satellite data, however, come with an important limitation—they are typically available only for aggregated geographic units such as grids or villages. So-called "bottom-up" statistical techniques combine survey data with remote sensing indicators to permit greater geospatial precision (box 4.4).

### Data profiling and discrimination

Because of the complexity and unstructured nature of private intent data, data scientists are increasingly relying on modern machine learning methods and algorithms to analyze them.[72] These algorithms can contain millions of parameters, which can be extremely costly and time-consuming to calibrate.[73] Machine learning experts thus typically rely on algorithms that are "pre-trained" using very large quantities of private intent data to make them easy to use for a variety of tasks. Although these algorithms are extremely useful for extracting insights from complex datasets, researchers in recent years realized that biases in the data used to calibrate these algorithms could contribute to discrimination,[74] with adverse consequences for people's welfare. Other research found that a machine learning tool created

## Box 4.4 Using statistical methods and private intent data to improve representativeness and geospatial precision

Combined data sources, by improving the representativeness and precision of survey data, enable indicators to be reported at finer spatial scales. One statistical approach to improving representativeness typically used when combining survey data with mobile phone data or satellite imagery is to average the data from different sources using a common geographic unit of analysis. For example, a welfare measure such as an asset index could be averaged across all households in a village (enumeration area). The results are then related to satellite imagery or mobile phone data. This procedure works well when extrapolating from imagery to predict average consumption for countries or large areas not covered by a survey. This method can also be used to generate local estimates of welfare within a country, provided that an appropriate statistical method is used to directly incorporate information from the sample into the estimation procedure to obtain more precise estimates.

Facebook engineers have used deep-learning algorithms to detect buildings in satellite imagery, allowing them to downscale population estimates from the census to a much finer spatial resolution. However, these methods have significant drawbacks. Predictions based on models specified at aggregate levels will generally not deliver precise estimates unless they are combined in an appropriate way with survey-based estimates. Precision is an important consideration because most national statistical offices will not publish imprecise estimates due to quality concerns. Furthermore, geographic downscaling relies on a few key assumptions. Facebook assumes the population is distributed in proportion to the "built-up area," which leads to inconsistencies between the estimates and the census. For example, because a smaller portion of buildings in urban areas are residential, relying on built-up area to distribute population could exaggerate population counts in urban areas compared with rural ones.

An alternative method for estimating the population of small areas is to use "bottom-up" methods that draw on data from survey listing exercises rather than "top-down" disaggregation of census data. "Bottom-up" techniques offer the important advantage of being able to produce updated population estimates without a census at a fraction of the cost. They use survey data to calibrate a model that relates population in the areas sampled by the survey to remote sensing indicators. Geospatial indicators that predict population density include the geographic size of the village, the number of buildings, the extent of built-up area, and the presence of nighttime lights. The model can then be used to generate population estimates nationwide. Similar methods can be used to generate more precise estimates of nonmonetary poverty.[a] They likely could be applied to a variety of socioeconomic indicators, including monetary poverty, labor market outcomes, health outcomes, and educational attainment.

a. Masaki et al. (2020).

to predict the future criminal behavior of defendants in the United States embedded racial discrimination: black defendants were twice as likely as their white counterparts to be falsely classified as future criminals.[75] Similarly, image search engines such as Flickr, which have been the source of training data for various computer vision algorithms, have been shown to overrepresent light-skinned men between the ages of 18 and 40, leading to poorer performance by these algorithms when making predictions of underrepresented categories such as women or minorities.[76]

Similar issues arise when machine learning algorithms are pre-trained using text containing racist and sexist stereotypes. Text generation algorithms trained on massive online text databases that were scraped from the web, such as the GPT-2 database created by Open AI, have been found to generate racist and anti-Semitic text in response to specific inputs.[77] When trained on Google News, word-embedding algorithms aimed at measuring the similarity between words tend to propagate the sexist biases reflected in the text, highlighting similarities between "man" and "computer programmer," whereas "woman" appears to be associated with "homemaker."[78] Arguably, such discrimination can have larger consequences in lower-income countries, which typically lack safety nets and social protection mechanisms.

### Data transparency and manipulation

Both the data-generating process and the algorithms used to process private intent data suffer from a lack of transparency. The algorithms used by search engines

are not public, and they are constantly optimized to improve users' experience. This process can lead to inaccurate predictions of policy outcomes, such as the notorious Google Flu Trends index. In 2009 a team of scientists at Google published a paper describing an innovative method to predict the number of flu cases in the United States using the volume of search terms related to the flu on Google.[79] Their Google Flu Trends index was initially able to predict official numbers ahead of the US Centers for Disease Control and Prevention (CDC), until it made headlines in 2013 for incorrectly predicting twice the number of actual flu cases. Scientists investigating what went wrong realized that many search terms used as predictors were associated with the onset of winter instead of the onset of colds.[80] This "overfitting" is a major concern when private intent datasets containing high-dimensional data (that is, data with a high number of features or independent variables) are used to nowcast policy outcomes that are infrequently observed. New generations of forecasting models based on private intent data should aim to rely on information coming from multiple private data sources to avoid being too dependent on the idiosyncrasies of a single source.

Even when accurate, predictive models are often so opaque that their predictions cannot be easily communicated to policy makers. Because machine learning is increasingly used to shape development policies, more research is needed to make complex algorithms transparent and interpretable, thereby increasing their legitimacy and ensuring they do not contribute to unequal outcomes. More research is also needed to understand trade-offs between interpretability and predictive performance. For example, researchers have estimated models using data to predict poverty from satellite imagery in both Sri Lanka and Uganda by focusing on objects in images that correlate with standards of living such as roads, buildings, and cars.[81] In each country, the interpretable model performed as well as commonly used black-box computer vision algorithms, indicating that model interpretability does not necessarily come at the cost of performance. The performance of image recognition algorithms may be constrained, however, because they are initially trained to detect a wide variety of objects using millions of images, which may not isolate the most important portions of the images for the specific purpose of predicting poverty.

An additional challenge of relying on algorithms to design policy is that they can be manipulated. People can change their behavior in response to algorithmic decision-making to trick the system and maximize their interests. For example, the nonprofit GiveDirectly facilitates direct cash transfers to poor households. As a proxy for poor living conditions, satellite imagery was initially used to target households with thatched roofs. When GiveDirectly's methods became common knowledge, some families pretended to live in a thatched structure near their home to qualify for the aid.[82] This concern about manipulation increasingly motivates the design of machine learning algorithms that assign more weight to personal characteristics less likely to be subject to manipulation.[83]

## Investments in data innovations: Building a culture of data

Effectively leveraging new types of data requires investing in human capital, data sharing, and research in lower-income countries. This section describes areas in which governments, donors, and advocates of corporate social responsibility can help promote innovative uses of data for public benefit in lower-income countries, thereby helping to build a culture for the use of data and evidence.

### Investing in people

*Building the skills of analysts and decision-makers.* Leveraging the comparative advantages of public intent and private intent data requires a long-term approach to enhancing domestic human capital in lower-income countries. Investments in human capital should focus on decision-makers and analysts. Strengthening the data and statistical literacy of decision-makers can help them understand the potential utility and limitations of these new data sources. This understanding is key for them to champion a data innovation agenda and advocate for the required human, technological, and financial resources. Analysts, on the other hand, need unique skill sets to leverage private intent data that bridge many disciplines, including statistics, economics, computer science, geographic information systems (GIS), and the multidisciplinary field of data science. Although many of these skills are akin to those needed to bolster the capacities of national statistical offices (NSOs), teams with exposure to private sector data and data systems will be able to work more efficiently across data types and foster collaboration.

At the country level, it is critical to build analysts' skills to integrate public intent and private intent data for public policy design and evaluation. These skills include *data engineering* to manage, process, and link public intent and private intent data; *analyzing* integrated datasets using traditional statistical and econometric methods and the latest advances in machine

learning; and *visualizing* the emerging insights. These skills must be augmented with acute awareness of the *ethics* and *data protection* dimensions of public intent and private intent data sources. Several competency frameworks developed for big data analytics are useful for a more granular understanding of the skill sets required for data acquisition, processing, analysis, visualization, and reporting.[84] These broad directions for capacity building focus on catalyzing the use of new data sources, in contrast to the recommendations presented in chapter 2, which focus on strengthening data production within the public sector.

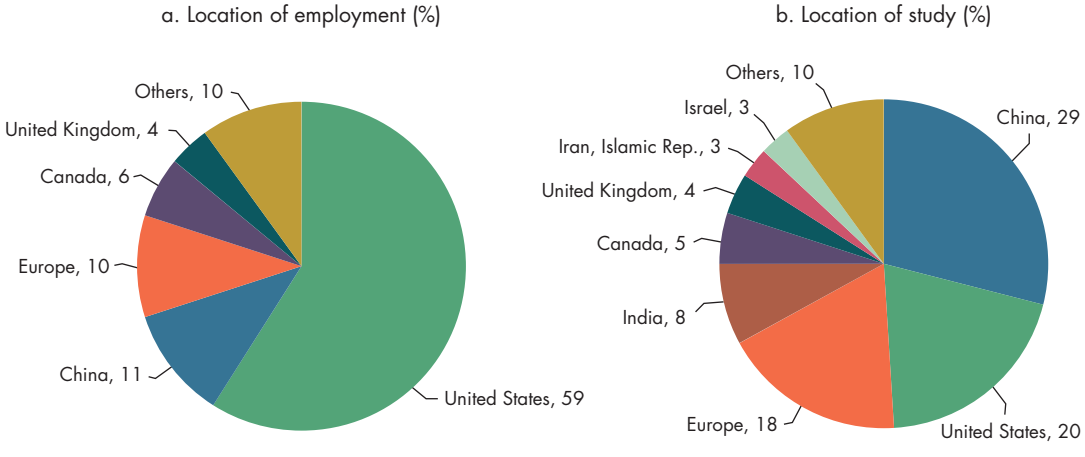
*Enhancing tertiary education.* The long-term process of acquiring these skills begins by enhancing tertiary education. Because of the wide array of competencies that data scientists are expected to possess, university and graduate degree programs may have to be altered, particularly in lower-income countries. Students need the foundational statistical skills central to understanding and using public intent data, as well as the frontier skills in artificial intelligence (AI) and machine learning at the heart of leveraging the value from the integration of public intent and private intent data. The curricula of degree programs—in the fields of statistics, economics, computer science, and GIS—could be revised to align formal education with the practical demands of jobs in data analytics. In addition, new degree, graduate, and certificate programs with a data science theme could be established.

*Promoting partnerships with universities and private companies in higher-income countries.* Such partnerships can be instrumental in achieving these education goals and enhancing training in contemporary data topics such as machine learning and AI. These types of initiatives can help tailor research in lower-income countries that leverages private intent data to local contexts and hires more local researchers. This would be a welcome trend because this research field has been predominantly led by principal investigators who are not nationals of these countries.

Proficiency with AI is one of the most coveted data skill sets. It involves feeding computers large amounts of data to train them to identify patterns and make predictions. For example, seismic activity data are crunched by computers to learn how to predict earthquakes,[85] and satellite images of agricultural areas are processed to estimate crop yields.[86] According to an analysis of self-reported job skills on the professional network platform LinkedIn, the United States leads in AI, followed by China (see figure 4.2).[87] Low- and middle-income countries need to catch up to these emerging trends in skills. In South Africa, the minister of communications and digital technologies argues there is no shortage of talent in the Africa region, but rather a lack of visionary policy makers to drive digitization and enable key infrastructure such as data centers and cloud computing.[88]

Technical training can sometimes be obtained cheaply or at no cost. Some digital companies provide free online training, and their certifications often attract job seekers.[89] Cisco's Networking Academy has trained more than 10 million people in low- and

**Figure 4.2 Artificial intelligence specialists gravitate to the US market, no matter where they are educated**




*Source:* MacroPolo, "The Global AI Talent Tracker," https://macropolo.org/digital-projects/the-global-ai-talent-tracker/. Adapted with permission of MacroPolo/Paulson Institute; further permission required for reuse. Data at http://bit.do/WDR2021-Fig-4_2.

*Note:* Country affiliations are based in panel a on the headquarters of institutions in which researchers currently work and in panel b on the country in which researchers received their undergraduate degree.

middle-income countries, often in partnership with local academic institutions with no or low-cost tuition. It also offers free online courses.[90] Although basic tech knowledge is needed to participate in these options, these offerings suggest that relevant training can be obtained in many developing countries at low cost provided good broadband internet connectivity is available. Popular cloud data management and analytical applications also feature graphical user interfaces, making it easier for those without advanced coding skills to use them.

*Increasing training, mentorship, and on-the-job training.* Improvements along the formal education supply chain can be augmented by on-the-job training efforts that target a broad coalition of data producers and users across the public sector, academia, and civil society. Increasing access to online training platforms (such as DataCamp and Coursera) and online degree and certificate programs, as well as free courses offered by prestigious universities in higher-income countries, can help build capacity across an impressive array of topics related to both foundational and frontier data analytics. These activities could be supplemented by continued support of emerging data science initiatives that provide scope for collaboration, mentorship, and learning, including the Deep Learning Indaba Institute,[91] Data Science Africa conferences,[92] and the competition platform Zindi.[93]

Skills training companies and platforms have recently surfaced supporting the development of digital data skills in developing countries and linking trainees to employers. Upskilling platforms such as Andela and Gebeya in Africa and Revelo in Brazil train students in data analytics and software development. Andela, founded in 2014, is training young people to meet the demand for information technology (IT) talent globally and within Africa. Gebeya, founded in 2016, matches trainees with companies in Africa.[94] And data labeling companies such as CloudFactory in Kenya and Nepal and Samasource in Kenya are creating jobs for cleaning, categorizing, and labeling data used for AI applications.[95]

As for its continued support of short-term training and mentorship programs in lower-income countries, the international community should evaluate the conditions for achieving sustained improvements in local capacity to identify short-term capacity-building models that hold promise.

*Strengthening data literacy among senior leadership and creating institutional environments that encourage the use of sophisticated data and evidence.* The big push to build an army of data scientists for jobs in the public sector, private sector, and civil society must be complemented with efforts to create enabling institutional and leadership environments (see chapter 8) that place a high premium on the use of data and evidence—both internally for management of these institutions and externally for understanding and producing policies that enhance welfare.

To help strengthen data literacy, especially in low-capacity settings, regional and international development partners can leverage their expertise or technical partnerships to provide governments with technical assistance. They can also organize objective peer reviews for gauging the relevance and accuracy of complex research that hinges on the integration of public intent and private intent data sources, including efforts sponsored by international agencies themselves (see spotlight 2.2).

On the whole, strengthening the data literacy of the senior leadership of public sector institutions will not guarantee that they will seek data and evidence when designing policies, especially if their insights do not appear to contribute to the political objectives of their government (see chapter 8). As discussed in chapter 2, mutually reinforcing constraints in financing, human capital, data governance, and data demand must be overcome as part of a long-term, holistic plan backed by domestic support from politicians of the major political parties, academia, and civil society.

In the short term, strengthening human capital in NSOs and line ministries in lower-income countries in the production and use of public intent data will indirectly contribute to the pool of skill sets required for public intent and private intent data to be integrated into official statistics and knowledge products generated within the public sector (see chapters 2 and 9 for further discussion). International organizations can provide these institutions with technical assistance to cultivate open data practices and to build skills in the creation and dissemination of public use census, survey, and administrative datasets that are subject to international best practices in deidentification. This effort can catalyze downstream research that brings together public intent and private intent data sources.

Statistical capacity-building projects financed by international organizations and traditionally focused on the production and use of public intent data should be expanded systematically to allow for investments in skills critical to the integration of public intent and private intent data sources. NSOs could establish a business line on *experimental statistics* (that is, statistics that leverage new data sources and methods to better respond to users' needs and can be viewed as official statistics "in the making"). This business line would provide a more direct route to investing in staff

who can conduct cutting-edge research grounded in synergies among public intent and private intent data sources.[96]

*Revamping NSOs to perform nontraditional roles with private intent data.* In general, for NSOs to maintain relevance in a landscape in which they no longer generate the majority of the data, they should be empowered data stewards endowed with qualified staff who can perform nontraditional roles. NSOs must be able to field requests for accessing confidential data that can be used to calibrate and validate models that fuse public intent and private intent data sources. By pursuing a work program on experimental statistics, NSOs should aspire to be proactive contributors to research that would assess the public intent data requirements of synergistic applications. The Data Science Campus in the United Kingdom's Office for National Statistics (ONS) is an example of a unit in an NSO that is tasked with leveraging the latest advances in data science and the synergies between public intent and private intent data sources to serve the public good. The Campus works on data science projects not only for the ONS, but also for the UK government as well as international organizations in collaboration with partners from academia and the private sector.[97] Twinning arrangements between the NSOs in high-income countries with similar initiatives and NSOs in low- and middle-income countries can be one way to strengthen NSO capabilities in low-capacity environments to create units akin to the ONS Data Science Campus.

NSOs will also need to grapple with data protection issues. They must, for example, determine whether spatially deidentified data are sufficient for calibration purposes and what minimum volume and scope of confidential data will have to be accessed for specific applications. NSOs also can carefully identify applications in which access to confidential data are not required. However, accommodating requests for applications with well-defined and well-articulated confidential data needs or responding to time-sensitive requests tied to immediate policy needs (such as a humanitarian or disaster response) ultimately require that NSOs have personnel who are trained in data protection and law and who can enter into and enforce data sharing agreements to mitigate data protection risks. To fulfill these roles, NSOs must receive a significant infusion of financial and human capital and should consider actively engaging—at least in the short term—international organizations or academic institutions and research organizations, at both the local and international levels, to bridge the gaps in internal institutional and technical capacity.

*Investing in data accessibility.* Accessing private intent data remains challenging, especially in lower-income countries. Large barriers, such as protecting customers and maintaining competitive advantages, prevent companies from sharing their data. In addition, pulling data from a company's database requires computing and human resources that are typically outside of a business's key performance indicators. If a public organization has poorly formulated requests for a company's data, compiling and exporting data can become a time-intensive burden on companies. Even if a company is willing and able to share its data, because of the diversity of private intent data types it is difficult to create standards to share data. Shared data must have clear documentation, be in a usable format that is interoperable with other private and public datasets for integration, and have been deidentified. Creating these types of standards may require third parties to coordinate efforts and will place more resource burdens on companies.

*Utilizing data collaboratives and research partnerships.* These cooperative arrangements are essential ways for different sectors, research institutions, and governments to share data. The Open Data Lab describes data collaboratives as moving beyond public-private partnerships to pool data resources that researchers use for public benefit.[98] A successful example can be found in Nairobi, Kenya, where researchers have partnered with local government agencies to develop spatially integrated road safety datasets with inputs from administrative, social media, private, and traditional sources (see spotlight 4.2 for details).

Data collaboratives can be coordinated by civil society or universities, or through corporate social responsibility programs. Facebook's Data for Good initiative is an example of how technology companies can be incentivized to share their data through corporate social responsibility programs. By leveraging customer data and engaging with civil society and university partners, Facebook is offering a suite of innovative datasets intended to aid public policy decisions. Offerings range from mobility data and downscaled population maps to data on electric grid coverage.

Data collaboratives and research partnerships could provide companies' application programming interfaces (APIs) and cloud services through tiers made available to the public sector. Social media platforms such as Twitter provide APIs so that users can download their text data using free tiers.

Private companies could be encouraged to share their data at reduced cost for public initiatives, with special grants for researchers or tax breaks for the

data provider. Cloud computing services, such as Google Cloud and Amazon Web Services, are offering small education grants to researchers to access the computing infrastructure needed to leverage these datasets, which are often large. Flowminder, a Swedish nongovernmental organization, provides code, instructions, and support for mobile network operators to aggregate, deidentify, and share their CDR data. Their open-source tool, FlowKit, provides APIs, code, and databases to aid companies sharing these sensitive records with researchers.[99]

Trusted intermediaries are building platforms that provide researchers with private intent data or facilitate sending programming code to private companies, which can, in turn, run the code with their private intent data on behalf of the researcher and share aggregated research insights. Opportunity Insights, a nonpartisan, nonprofit research organization based at Harvard University, offers a Track the Recovery platform that gives researchers access to near real-time economic data to understand the COVID-19 policy response in the United States. As the broker of the data sharing agreements, Opportunity Insights deidentifies data to facilitate sharing by protecting customers' and companies' data. For example, they protect companies' data through aggregation and by creating relative indicators that mask actual revenue and profit. OPAL ("Open Algorithms") takes a different approach. OPAL is a nonprofit partnership created by groups at MIT Media Lab, Imperial College London, the financial company Orange, the World Economic Forum, and the Data-Pop Alliance. Its platform allows researchers to send companies certified open-source algorithms that are then run behind the companies' firewalls.

Despite the promise of these innovative data-sharing pathways, many are not available in lower-income countries. For example, these countries rarely participate in data collaboratives, according to data compiled by the Open Data Lab.[100] A similar trend can be seen in the limited number of studies on lower-income countries that leveraged private intent data in the early stages of the COVID-19 pandemic (box 4.2). More investments are needed in accessing private intent data for public benefit in lower-income countries.

### Investing in research

Investments in research are needed to develop methods and enable lower-income countries to grow research programs that leverage private intent data for public policy. The research community can achieve quick wins by focusing on foundational areas such as testing whether validated methods in one region translate to contexts where data are sparse. Over the longer term, research strategies would benefit from building validation and training datasets in lower-income countries to avoid issues similar to data profiling and discrimination when using pre-trained models from higher-income countries. The selections that follow describe some of the high-priority research needed to advance the use of private intent data for public benefit in the short and long term.

*Shorter-term research needs.* Because much of the current innovation in using private intent data is led by researchers and technology companies in higher-income countries, many of the available methods are not tailored to the development context.[101] Even when a solution is developed for and validated in a particular lower-income country, understanding whether and when the solution can be extended to other lower-income countries can enable research in data-sparse contexts. For example, even though international phone call usage correlates with wealth more strongly in Rwanda than in Balkh province in Afghanistan, such a finding can still be useful in contexts such as Balkh province.[102] Similarly, granular poverty maps that use digital trace data from mobile phones hold great potential for better targeting social services, but the patterns that algorithms use to make poverty predictions may differ from context to context.[103] Research is needed to determine when granular poverty estimates created for one country can be transferred to another country and when they will lead to misleading maps.

In the short term, researchers also need to produce methods that preserve privacy while combining public intent and private intent data. As more datasets are made available to researchers and decision-makers, more opportunities arise to reverse-engineer traditional deidentification methods. If these risks are not eliminated, individuals and companies may be reluctant to share their data. One example of how public intent data are being designed to prevent de-anonymization is the GPS data collected from household surveys for the Demographic and Health Survey and the Living Standards Measurement Study. Even if surveys collect GPS-based locations for communities and households, the resulting data are not included in public use datasets to ensure the confidentiality of respondents. Any third-party user that obtains DHS- or LSMS-type survey data has access only to spatially offset locations of survey enumeration areas. For example, a household's location is represented using the 10 square kilometer area

within which the house is located. These surveys are also good examples of providing documentation that makes the precision and accuracy of the deidentified data explicit.

Concerns about data protection have limited the eagerness to share data, even in critical times such as during the Ebola crisis. Historically, data deidentification techniques have maintained equilibrium between the producers and consumers of data, preserving individuals' privacy while limiting information loss. However, deidentification techniques have proven to be increasingly imperfect with high-dimensional private intent data. Despite the use of standard deidentification techniques, one study found that four data points were enough to reidentify 95 percent of individuals in a mobile phone dataset of 1.5 million people.[104] In this context, new data sharing frameworks have been proposed to mitigate privacy risks while maximizing the informative potential of private intent data. Researchers have proposed four models for use of mobile phone data, depending on the level of risk tolerance and the number of potential third-party users.[105] The Social Science One initiative, which allows researchers from academic institutions to access Facebook data at scale,[106] is an example of how new data sharing frameworks could be applied to access private intent data, paving the way for future public-private collaborations. As noted, more research will be needed to design methods that allow the privacy of private intent data to be protected, while minimizing the loss of precision associated with using these data in applications aimed to inform public policy.[107]

In a context of low data and coding literacy, off-the-shelf programming tools can lead to more effective and responsible use of private intent data. Flowminder's FlowKit is an example of an open-source solution that helps companies to deidentify, clean, and export their data effectively for policy applications.[108] Using FlowKit, Flowminder and its partners have been able to rapidly integrate CDRs into the COVID-19 response. Aequitas is another open-source toolkit that provides an intuitive way to audit machine learning models for discrimination and bias.[109] These types of tools enable researchers to access data and companies to share data without the need for specialized skills to collate and deidentify the datasets. Ideally, these research tools should be designed to promote access to data and to share technical knowledge between lower-income countries, from higher-income countries to lower-income countries, or from lower-income countries to higher-income countries. Research funding would not only operate on short-term project cycles but also support the institutional setup of research labs and institutes in lower-income countries.

*Longer-term research needs.* These needs include devising best practices and quality standards. Most decision-makers will not be well versed in the latest data methods. Best practices and quality standards can facilitate trust in leveraging new data types for policy. These types of standards and governing institutions are available for public intent data. Conceivably, then, they could be translated for private intent data. For example, traditional data collection using sample surveys has many imperfections, but by studying them extensively, the research community has come up with ways to address them or quantify the errors they introduce.[110] In the same way, researchers need to study the limitations of private intent data and develop the appropriate quality standards for their use in public policy. For example, there is currently no consensus on the criteria needed to determine whether a poverty map is fit for use in resource allocations.

International organizations can play a major role in this process by providing platforms for discussing, formulating, and promoting these practices and standards (see spotlight 8.1). The working groups established under the auspices of the United Nations Statistical Commission on household surveys, open data, and big data may provide insight into the types of commissions that could be established.[111]

### Coordinating investment

In the longer term, coordinated investment in high-quality training data from lower-income countries will also be needed. Currently, private intent data are mostly repurposed for machine learning applications, which require high-quality data collected on location via remote sensing to train algorithms. For example, over the last five years pioneering research on small-scale farming systems has successfully combined high-quality georeferenced survey data with high-resolution, multispectral satellite imagery from public sources (Sentinel-2) and private sources (Planet and Maxar, formerly Digital Globe) to obtain crop yield estimates on individual plots.[112] These efforts have shown the importance of using high-quality ground data—including georeferenced plot outlines and objective measures of crop yields—to calibrate and validate remote sensing models that can, in turn, churn out high-resolution grids for crop types and crop yields for entire regions and countries.

One of the challenges preventing the rapid scale up of these efforts is the lack of knowledge on the required volume and content of georeferenced

microdata that should be collected through surveys to inform downstream remote sensing applications capable of meeting needs for spatially disaggregated estimation and reporting. These challenges could be addressed by research. Similarly, to analyze natural language data to, for example, measure attitudes on certain topics in the population, researchers typically rely on pre-trained language models (that is, models already trained on a large corpus of text). The lack of these pre-trained language models in languages other than the major ones has been a barrier to the analysis of text data in low- and middle-income countries.[113] If the people in these countries are themselves the designers, curators, and owners of location-specific, high-quality training data to test private intent data, the center of research gravity would naturally shift toward lower-income countries.

Table 4.1 summarizes selected short- and long-term research needs. Answers to the questions listed in the table will vary in accordance with the development

**Table 4.1 Selected research gaps to be addressed to advance the use of private intent data for development**

| Research area | Examples of research gaps |
|---|---|
| Societal impacts | • How do we ensure that algorithmic-based policy making can lead to fair outcomes?<br>• How can we increase the transparency and interpretability of policy predictions using private intent data?<br>• How can we design algorithms that can be safeguarded against manipulation?<br>• What are the trade-offs between granularity and precision, and what is the optimal mix for targeting of development programs? |
| Quality standards | • How can standards be created, agreed on, updated, and communicated to the general development community? Who needs to be part of these conversations?<br>• To ensure that policy makers can trust and use results, what should the standards be for accuracy and precision for frontier applications that use private intent data or that combine public intent and private intent data? |
| External validity[a] | • How promising is the approach of building models in countries that have data and applying them to countries with limited data?<br>• How can issues akin to data profiling and discrimination be avoided when using pre-trained models from higher-income countries in cases of novel development use?<br>• To what extent can applications that combine public intent survey data with private intent data predict values calculated from census data within a country? |
| Machine learning | • How does the approach to machine learning and spatial feature selection need to change from common machine learning tasks to more specialized tasks that will aid development policy?<br>• Which features best predict spatial variation in development outcomes in different contexts? What are the trade-offs between predictive accuracy and cost? |
| Training and validation data | • What should be the required volume of and approach to public intent data collection for calibrating and validating machine learning algorithms that combine public intent and private intent data? |
| Deidentification[b] | • How do deidentification methods need to change to protect individuals and companies when private intent data are used for public benefit?<br>• How does (spatial) deidentification of public intent data affect the accuracy and precision of applications that use public intent data to calibrate and validate machine learning algorithms that combine public intent and private intent data? |
| Capturing longitudinal change | • How do accuracy and precision differ in applications that aim to estimate longitudinal change versus obtaining cross-sectional predictions for the same development outcome?<br>• What features best predict longitudinal change in different contexts?<br>• How can we ensure the stability over time of algorithms aimed at predicting changes in policy outcomes?<br>• When public intent survey data are combined with imagery—specifically, spatial features (predictors) extracted via deep-learning techniques—in order to derive high-resolution estimates of a development outcome, how do the spatial, spectral,[c] and temporal resolution of satellite imagery affect the accuracy and precision of the predictions for the outcome of interest?<br>• Do these effects vary based on the decisions on the size of satellite imagery grids that are processed for extracting spatial features? |

a. External validity relates to the research findings of one location holding true in another location.
b. The term deidentification is used instead of anonymization because, although data are processed to deidentify any individual, these data may become identifiable in the future as computing and machine learning advance. Thus data may never be truly anonymized.
c. "Spectral" refers to different wavelengths on the visual spectrum. Satellite images typically have multiple "bands" that capture different spectral ranges.

outcome/process that researchers are aiming to better measure and understand through the use and augmentation of private intent data. For example, the requirements for high-resolution estimation of population density will differ from requirements for estimating crop yields.

The growing availability and use of private intent data for development purposes have potentially large benefits, especially when paired with public intent data. However, the way forward requires a conducive and enabling environment that trains both analysts and higher-level decision-makers to consider critically issues of data protection, discrimination, manipulation, representativeness, and transparency. Repurposing and combining public intent and private intent data are central to getting more value from data, but the benefits must be shared equitably while safeguarding against harmful outcomes. Part II of this Report describes the building blocks of a social contract that enables such data flows, including infrastructure policies, legal and regulatory frameworks for data, related economic policies, and the institutions of data governance.

## Notes

1. Bengtsson et al. (2011).
2. Chetty et al. (2020); Oliver et al. (2020).
3. Beraja, Yang, and Yuchtman (2020).
4. Salganik (2017).
5. Salganik (2017).
6. Serajuddin et al. (2015).
7. IEAG (2014).
8. Demombynes and Sandefur (2015).
9. Tiecke and Gros (2016).
10. Stephens-Davidowitz (2017).
11. For the 2008 and 2012 US presidential elections, Stephens-Davidowitz (2017) found that an area's search rate for terms with racial overtones was a robust negative predictor of presidential candidate Barack Obama's vote share.
12. WHO (2008).
13. Adda (2016).
14. Ihantamalala et al. (2018); Milusheva (2020); Wesolowski et al. (2012).
15. González, Hidalgo, and Barabási (2008); Le Menach et al. (2011); Tatem et al. (2009).
16. Wesolowski et al. (2012).
17. Wesolowski et al. (2012).
18. Peak et al. (2018). After the outbreak, they studied how mobile phone data for Sierra Leone could have been used to evaluate the impacts of interventions meant to decrease travel during the epidemic.
19. COVID-19 National Emergency Response Center (2020).
20. Burns (2020).
21. Chang et al. (2020); Maas et al. (2019).
22. Aktay et al. (2020).
23. Lai et al. (2020); Pepe et al. (2020).
24. Salathé et al. (2012).
25. McCall (2020).
26. PAHO and WHO (2016).
27. McGough et al. (2017).
28. Kraemer et al. (2019).
29. Yang et al. (2017).
30. Milinovich et al. (2014).
31. Internal Displacement Monitoring Center (IDMC), Data of GIDD (Global Internal Displacement Database), https://www.internal-displacement.org/database/displacement-data.
32. Ritchie and Roser (2019).
33. BBC News (2018); CNN Indonesia (2018).
34. Bengtsson et al. (2011); Lu, Bengtsson, and Holme (2012); Wilson et al. (2016).
35. Robinson, Power, and Cameron (2013).
36. Robinson, Power, and Cameron (2013).
37. Kongthon et al. (2012).
38. Resch, Usländer, and Havas (2018).
39. Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude toward a topic or product is positive, negative, or neutral. See "sentiment analysis," Lexico, Oxford University Press, https://www.lexico.com/en/definition/sentiment_analysis.
40. Reynard and Shirgaokar (2019).
41. See "Case Study 5: Delivering Remote Flood Analytics as a Scalable Service," pages 61–68 in Sylvester (2019).
42. Blumenstock, Cadamuro, and On (2015); Jean et al. (2016); Yeh et al. (2020).
43. Blumenstock, Cadamuro, and On (2015).
44. Frias-Martinez, Frias-Martinez, and Oliver (2010).
45. Aiken et al. (2020).
46. Llorente et al. (2015).
47. Glaeser, Kim, and Luca (2018).
48. Blumenstock (2016).
49. Bonnet, Lechat, and Ridde (2018).
50. Williams, Idowu, and Olonade (2015).
51. Kelley, Lane, and Schönholzer (2020).
52. Dai and Sujon (2019).
53. Milusheva et al. (2020).
54. Kelley, Lane, and Schönholzer (2020).
55. Pratihast et al. (2014).
56. See World Resources Institute, Global Forest Watch (dashboard), https://www.globalforestwatch.org/.
57. WRI (2019).
58. See, for example, Janaagraha Centre for Citizenship and Democracy, I Paid a Bribe (dashboard), https://www.ipaidabribe.com/about-us#gsc.tab=0.I.
59. Hlatshwayo et al. (2018).
60. Inter-American Development Bank, "Countries That Have Already Implemented the Investment Map Initiative," https://www.iadb.org/en/reform-modernization-state/countries.
61. Marshall (2012).
62. Deaton (2008); Falk et al. (2018). For a list of projects that have used Gallup World Poll data, see Gallup, "Working Together to Change the World," https://www.gallup

.com/analytics/318176/public-sector-success-stories.aspx.
63. Goldstein, Gonzalez Martinez, and Papineni (2019).
64. Goldstein, Gonzalez Martinez, and Papineni (2019).
65. See Google, Community Mobility Reports (database), https://www.google.com/covid19/mobility/?hl=en.
66. See Ookla, Speedtest Global Index (database), https://www.speedtest.net/global-index.
67. Davis et al. (2010).
68. Mehrabi et al. (2020).
69. Blumenstock and Eagle (2012).
70. Wesolowski et al. (2012).
71. Frankfurter et al. (2020).
72. Jean et al. (2016).
73. Strubell, Ganesh, and McCallum (2019).
74. Zou and Schiebinger (2018).
75. Angwin et al. (2016).
76. Buolamwini and Gebru (2018).
77. Wallace et al. (2019).
78. Bolukbasi et al. (2016).
79. Ginsberg et al. (2009).
80. Lazer et al. (2014).
81. Ayush et al. (2020); Engstrom, Hersh, and Newhouse (2017).
82. Blumenstock (2018).
83. Björkegren, Blumenstock, and Knight (2020).
84. Carretero, Vuorikari, and Punie (2017); GSS (2016); Vale and Gjaltema (2020).
85. Perol, Gharbi, and Denolle (2018).
86. RTI International, "Impact: Using Satellite Images and Artificial Intelligence to Improve Agricultural Resilience," https://www.rti.org/impact/using-satellite-images-and-artificial-intelligence-improve-agricultural-resilience.
87. Perisic (2018).
88. ITU (2020).
89. Flowers (2019).
90. Cisco Systems, "Cisco Networking Academy," https://www.cisco.com/c/en/us/about/csr/impact/education/networking-academy.html.
91. Deep Learning Indaba Institute, https://deeplearningindaba.com/2020/.
92. Data Science Africa, http://www.datascienceafrica.org/.
93. Zindi (2020).
94. Buckholtz (2019).
95. Kaye (2019).
96. See, for example, the related efforts under the European Statistical System by Eurostat, Statistics Denmark, Destatis (Germany), National Statistics Institute (Spain), ISTAT (Italy), Central Statistical Bureau of Latvia, Statistics Netherlands, Statistics Poland, Statistics Portugal, National Institute of Statistics (Romania), Statistics Finland, Statistics Iceland, and the Federal Statistical Office (Switzerland)—see European Statistical System, Eurostat, "Experimental Statistics," Luxembourg, https://ec.europa.eu/eurostat/web/ess/experimental-statistics.
97. For more information on the ONS Data Science Campus and its projects, see Data Science Campus, Office for National Statistics, "Data Science for Public Good: Projects," https://datasciencecampus.ons.gov.uk/projects/.
98. GovLab, Tandon School of Engineering, New York University, "Data Collaboratives," https://datacollaboratives.org/.
99. Flowminder Foundation, "FlowKit CDR Analytics Toolkit," https://flowkit.xyz/.
100. GovLab, Tandon School of Engineering, New York University, "Data Collaboratives," https://datacollaboratives.org/.
101. Blumenstock (2018).
102. Aiken et al. (2020).
103. Blumenstock (2018).
104. de Montjoye et al. (2013).
105. de Montjoye et al. (2013).
106. Social Science One, Institute for Quantitative Social Science, Harvard University, "Building Industry-Academic Partnerships," https://socialscience.one/home.
107. Dwork and Roth (2014).
108. Flowminder Foundation, "FlowKit CDR Analytics Toolkit," https://flowkit.xyz/.
109. Saleiro et al. (2019).
110. Bethelehem (2009).
111. United Nations Statistical Commission, Statistics Division, Department of Economic and Social Affairs, United Nations, "Active Groups under the Statistical Commission by Pillar and Type of Group," https://unstats.un.org/unsd/statcom/groups/.
112. Burke and Lobell (2017); Gourlay, Kilic, and Lobell (2019); Jain et al. (2016); Lambert et al. (2018); Lobell et al. (2020).
113. Zindi (2020).

## References


Adda, Jérôme. 2016. "Economic Activity and the Spread of Viral Diseases: Evidence from High Frequency Data." *Quarterly Journal of Economics* 131 (2): 891–941.

Aiken, Emily L., Guadalupe Bedoya, Aidan Coville, and Joshua Evan Blumenstock. 2020. "Targeting Development Aid with Machine Learning and Mobile Phone Data: Evidence from an Anti-Poverty Intervention in Afghanistan." In *COMPASS '20: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 310–11. New York: Association for Computing Machinery.

Aktay, Ahmet, Shailesh Bavadekar, Gwen Cossoul, John Davis, Damien Desfontaines, Alex Fabrikant, Evgeniy Gabrilovich, et al. 2020. "Google COVID-19 Community Mobility Reports: Anonymization Process Description (Version 1.0)." April 8, 2020. https://arxiv.org/abs/2004.04145v1.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used across the Country to Predict Future Criminals, and It's Biased against Blacks." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ayush, Kumar, Burak Uzkent, Marshall Burke, David B. Lobell, and Stefano Ermon. 2020. "Generating

Interpretable Poverty Maps Using Object Detection in Satellite Images." Cornell University, Ithaca, NY. http://arxiv.org/abs/2002.01612.

BBC News. 2018. "Indonesia Earthquake and Tsunami: How Warning System Failed the Victims." *BBC News*, October 1, 2018. https://www.bbc.com/news/world-asia-45663054.

Bengtsson, Linus, Xin Lu, Anna Thorson, Richard Garfield, and Johan von Schreeb. 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): e1001083. https://doi.org/10.1371/journal.pmed.1001083.

Beraja, Martin, David Y. Yang, and Noam Yuchtman. 2020. "Data-Intensive Innovation and the State: Evidence from AI Firms in China." NBER Working Paper 27723, National Bureau of Economic Research, Cambridge, MA. https://www.nber.org/papers/w27723.

Bethlehem, Jelke. 2009. "The Rise of Survey Sampling." Discussion Paper 09015, Statistics Netherlands, The Hague.

Björkegren, Daniel, Joshua Evan Blumenstock, and Samsun Knight. 2020. "Manipulation-Proof Machine Learning." Cornell University, Ithaca, NY. http://arxiv.org/abs/2004.03865.

Blumenstock, Joshua Evan. 2016. "Fighting Poverty with Data." *Science* 353 (6301): 753–54. https://doi.org/10.1126/science.aah5217.

Blumenstock, Joshua Evan. 2018. "Don't Forget People in the Use of Big Data for Development." *Nature* 561 (7722): 170–72. https://doi.org/10.1038/d41586-018-06215-5.

Blumenstock, Joshua Evan, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264): 1073–76. https://doi.org/10.1126/science.aac4420.

Blumenstock, Joshua Evan, and Nathan Eagle. 2012. "Divided We Call: Disparities in Access and Use of Mobile Phones in Rwanda." *Information Technologies and International Development* 8 (2): 1–16.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." July 21, Cornell University, Ithaca, NY. https://arxiv.org/abs/1607.06520.

Bonnet, Emmanuel, Lucie Lechat, and Valéry Ridde. 2018. "What Interventions Are Required to Reduce Road Traffic Injuries in Africa? A Scoping Review of the Literature." *PLoS ONE* 13 (11): e0208195. https://doi.org/10.1371/journal.pone.0208195.

Buckee, Caroline O., Satchit Balsari, Jennifer Chan, Mercè Crosas, Francesca Dominici, Urs Gasser, Yonatan H. Grad, et al. 2020. "Aggregated Mobility Data Could Help Fight COVID-19." *Science* 368 (6487): 145–46. https://doi.org/10.1126/science.abb8021.

Buckholtz, Alison. 2019. "Africa's IT Talent Pool." *IFC Insights* (blog), December 2019. https://www.ifc.org/wps/wcm/connect/news_ext_content/ifc_external_corporate_site/news+and+events/news/insights/africa-it-talent.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *PMLR, Proceedings of Machine Learning Research*, vol. 81, *FAT 2018, Conference on Fairness, Accountability, and Transparency, 23–24 February 2018, New York, NY, USA*, edited by Sorelle A. Friedler and Christo Wilson, 77–91. Cambridge, MA: MIT Press. https://dam-prod.media.mit.edu/x/2018/02/06/Gender%20Shades%20Intersectional%20Accuracy%20Disparities.pdf.

Burke, Marshall, and David B. Lobell. 2017. "Satellite-Based Assessment of Yield Variation and Its Determinants in Smallholder African Systems." *PNAS, Proceedings of the National Academy of Sciences* 114 (9): 2189–94. https://doi.org/10.1073/pnas.1616919114.

Burns, Sarah. 2020. "How Anonymized Mobile Data Are Helping Ghana Fight COVID-19." Global Partnership for Sustainable Development Data, United Nations, New York. https://www.data4sdgs.org/news/how-anonymized-mobile-data-are-helping-ghana-fight-covid-19.

Carretero, Stephanie, Riina Vuorikari, and Yves Punie. 2017. "DigComp 2.1: The Digital Competence Framework for Citizens, with Eight Proficiency Levels and Examples of Use." JRC Working Paper JRC106281, Joint Research Center, EU Science Hub, Seville, Spain.

Chang, Meng-Chun, Rebecca Kahn, Yu-An Li, Cheng-Sheng Lee, Caroline O. Buckee, and Hsiao-Han Chang. 2020. "Modeling the Impact of Human Mobility and Travel Restrictions on the Potential Spread of SARS-CoV-2 in Taiwan." *medRxiv*, April 11, 2020. https://doi.org/10.1101/2020.04.07.20053439.

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Michael Stepner, and the Opportunity Insights Team. 2020. "How Did COVID-19 and Stabilization Policies Affect Spending and Employment? A New Real-Time Economic Tracker Based on Private Sector Data." NBER Working Paper 27431, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w27431.

CNN Indonesia. 2018. "BNPB: Seluruh Buoy Deteksi Tsunami di Indonesia Rusak." *CNN Indonesia*, September 30, 2018. https://www.cnnindonesia.com/nasional/20180930160115-20-334439/bnpb-seluruh-buoy-deteksi-tsunami-di-indonesia-rusak.

COVID-19 National Emergency Response Center. 2020. "Contact Transmission of COVID-19 in South Korea: Novel Investigation Techniques for Tracing Contacts." *Osong Public Health and Research Perspectives* 11 (1): 60–63. COVID-19 National Emergency Response Center, Epidemiology and Case Management Team, Korea Centers for Disease Control and Prevention, Cheongju, Republic of Korea. https://doi.org/10.24171/j.phrp.2020.11.1.09.

Dai, Fei, and Mohhammad Sujon. 2019. "Measuring Current Traffic Safety Culture via Social Media Mining." WTSC Report 2019-AG-2856, Washington Traffic Safety Commission, Olympia, WA. http://wtsc.wa.gov/wp-content/uploads/dlm_uploads/2019/10/Measuring-Traffic-Safety-Culture-via-Social-Media-Mining_Oct2019-1.pdf.

Davis, Kristin E., Burton Swanson, David Amudavi, Daniel Ayalew Mekonnen, Aaron Flohrs, Jens Riese, Chloe Lamb, and Elias Zerfu. 2010. "In-Depth Assessment of the Public Agricultural Extension System of Ethiopia and Recommendations for Improvement." IFPRI Discussion Paper 01041, International Food Policy Research Institute,

Washington, DC. https://www.ifpri.org/publication/depth-assessment-public-agricultural-extension-system-ethiopia-and-recommendations.

Deaton, Angus S. 2008. "Income, Health, and Well-Being around the World: Evidence from the Gallup World Poll." *Journal of Economic Perspectives* 22 (2): 53–72. https://doi.org/10.1257/jep.22.2.53.

Demombynes, Gabriel, and Justin Sandefur. 2015. "Costing a Data Revolution." *World Economics* 16 (3): 99–112.

de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. "Unique in the Crowd: The Privacy Bounds of Human Mobility." *Scientific Reports* 3 (1): 1376. https://doi.org/10.1038/srep01376.

Dureuil, Manuel, Kristina Boerder, Kirsti A. Burnett, Rainer Froese, and Boris Worm. 2018. "Elevated Trawling inside Protected Areas Undermines Conservation Outcomes in a Global Fishing Hot Spot." *Science* 362 (6421): 1403–07. https://doi.org/10.1126/science.aau0561.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends in Theoretical Computer Science* 9 (3–4): 211–407. http://dx.doi.org/10.1561/0400000042.

Engstrom, Ryan, Jonathan Samuel Hersh, and David Locke Newhouse. 2017. "Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being." Policy Research Working Paper 8284, World Bank, Washington, DC.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *Quarterly Journal of Economics* 133 (4): 1645–92. https://doi.org/10.1093/qje/qjy013.

Flowers, Andrew. 2019. "Indeed Tech Skills Explorer: Fastest-Rising Tech Skills." *Occupation Spotlight* (blog), November 26, 2019. https://www.hiringlab.org/2019/11/26/fastest-rising-tech-skills/.

Fraiberger, Samuel P., Pablo Astudillo, Lorenzo Candeago, Alex Chunet, Nicholas K. W. Jones, Maham Faisal Khan, Bruno Lepri, et al. 2020. "Uncovering Socioeconomic Gaps in Mobility Reduction during the COVID-19 Pandemic Using Location Data." Cornell University, Ithaca, NY. http://arxiv.org/abs/2006.15195.

Frankfurter, Zoe, Klaudia Kokoszka, David Locke Newhouse, Ani Rudra Silwal, and Siwei Tian. 2020. "Measuring Internet Access in Sub-Saharan Africa (SSA)." *Poverty and Equity Notes* 31 (August), World Bank, Washington, DC. https://openknowledge.worldbank.org/bitstream/handle/10986/34302/Measuring-Internet-in-Access-in-Sub-Saharan-Africa-SSA.pdf?sequence=1.

Frias-Martinez, Vanessa, Enrique Frias-Martinez, and Nuria Oliver. 2010. "A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records." In *Artificial Intelligence for Development: Papers from the AAAI Spring Symposium*, edited by Association for the Advancement of Artificial Intelligence, 37–42. Technical Report SS-10-01. Menlo Park, CA: AAAI Press.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (February): 1012–14. https://www.nature.com/articles/nature07634.

Glaeser, Edward L., Hyunjin Kim, and Michael Luca. 2018. "Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change." *AEA Papers and Proceedings* 108 (May): 77–82.

Goldstein, Markus P., Paula Gonzalez Martinez, and Sreelakshmi Papineni. 2019. "Tackling the Global Profitarchy: Gender and the Choice of Business Sector." Policy Research Working Paper 8865, World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/31747.

González, Marta C., César A. Hidalgo, and Albert-László Barabási. 2008. "Understanding Individual Human Mobility Patterns." *Nature* 453 (7196): 779–82.

Gourlay, Sydney, Talip Kilic, and David B. Lobell. 2019. "A New Spin on an Old Debate: Errors in Farmer-Reported Production and Their Implications for Inverse Scale–Productivity Relationship in Uganda." *Journal of Development Economics* 141 (November): 102376. https://doi.org/10.1016/j.jdeveco.2019.102376.

GSS (Government Statistical Service, UK). 2016. "Competency Framework for the Government Statistician Group (GSG)." GSS, Office of National Statistics, London.

Hlatshwayo, Sandile, Anne Oeking, Manuk Ghazanchyan, David Corvino, Ananya Shukla, and Lamin Leigh. 2018. "The Measurement and Macro-Relevance of Corruption: A Big Data Approach." IMF Working Paper WP/18/195, International Monetary Fund, Washington, DC. http://dx.doi.org/10.5089/9781484373095.001.

IEAG (Independent Expert Advisory Group on a Data Revolution for Sustainable Development). 2014. "A World That Counts: Mobilising the Data Revolution for Sustainable Development." Data Revolution Group, United Nations, New York.

Ihantamalala, Felana Angella, Vincent Herbreteau, Feno M. J. Rakotoarimanana, Jean Marius Rakotondramanga, Simon Cauchemez, Bienvenue Rahoilijaona, Gwenaëlle Pennober, et al. 2018. "Estimating Sources and Sinks of Malaria Parasites in Madagascar." *Nature Communications* 9 (1): 3897.

ITU (International Telecommunication Union). 2020. "Africa Is at the AI Innovation Table and 'Ready for the Next Wave.'" *ITU News*, June 23, 2020. https://www.itu.int/en/myitu/News/2020/06/23/07/55/AI-for-Good-2020-Africa-innovation.

Jain, Meha, Amit Srivastava, Balwinder Singh, Rajiv Joon, Andrew Mcdonald, Keitasha Royal, Madeline Lisaius, et al. 2016. "Mapping Smallholder Wheat Yields and Sowing Dates Using Micro-Satellite Data." *Remote Sensing* 8 (November): 860. https://doi.org/10.3390/rs8100860.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94. https://doi.org/10.1126/science.aaf7894.

Kaye, Kate. 2019. "These Companies Claim to Provide 'Fair-Trade' Data Work: Do They?" *MIT Technology Review*, August 7. https://www.technologyreview.com/2019/08/07/133845/cloudfactory-ddd-samasource-imerit-impact-sourcing-companies-for-data-annotation//.

Kelley, Erin, Gregory Lane, and David Schönholzer. 2020. "Monitoring in Target Contracts: Theory and Experiment in Kenyan Public Transit." Paper presented at Virtual BREAD/CEPR/STICERD/TCD Conference on Development Economics, October 1–3, 2020. https://youtu.be/TU-_xDR3x7I.

Klein, Brennan, Timothy LaRock, Stefan McCabe, Leo Torres, Filippo Privitera, Lake Brennan, Moritz U. G. Kraemer, et al. 2020. "Assessing Changes in Commuting and Individual Mobility in Major Metropolitan Areas in the United States during the COVID-19 Outbreak." Network Science Institute, Northeastern University, Boston. https://www.networkscienceinstitute.org/publications/assessing-changes-in-commuting-and-individual-mobility-in-major-metropolitan-areas-in-the-united-states-during-the-covid-19-outbreak.

Kongthon, Alisa, Choochart Haruechaiyasak, Jaruwat Pailai, and Sarawoot Kongyoung. 2012. "The Role of Twitter during a Natural Disaster: Case Study of 2011 Thai Flood." In *2012 Proceedings of PICMET '12: Technology Management for Emerging Technologies*, edited by Institute of Electrical and Electronics Engineers, 2227–32. Red Hook, NY: Curran Associates.

Kraemer, Moritz U. G., Nick Golding, Dionisio Bisanzio, Samir Bhatt, David M. Pigott, S. E. Ray, O. J. Brady, et al. 2019. "Utilizing General Human Movement Models to Predict the Spread of Emerging Infectious Diseases in Resource Poor Settings." *Scientific Reports* 9 (March): 5151. https://doi.org/10.1038/s41598-019-41192-3.

Lai, Shengjie, Nick W. Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R. Floyd, Amy Wesolowski, et al. 2020. "Effect of Non-Pharmaceutical Interventions to Contain COVID-19 in China." *Nature* 585 (7825): 410–13. https://doi.org/10.1038/s41586-020-2293-x.

Lambert, Marie-Julie, Pierre C. Sibiry Traoré, Xavier Blaes, Philippe Baret, and Pierre Defourny. 2018. "Estimating Smallholder Crops Production at Village Level from Sentinel-2 Time Series in Mali's Cotton Belt." *Remote Sensing of Environment* 216 (October): 647–57. https://doi.org/10.1016/j.rse.2018.06.036.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–05. https://doi.org/10.1126/science.1248506.

Le Menach, Arnaud, Andrew J. Tatem, Justin M. Cohen, Simon I. Hay, Heather Randell, Anand P. Patil, and David L. Smith. 2011. "Travel Risk, Malaria Importation, and Malaria Transmission in Zanzibar." *Scientific Reports* 1: 93. https://www.nature.com/articles/srep00093.

Llorente, Alejandro, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. "Social Media Fingerprints of Unemployment." *PLoS ONE* 10 (5): e0128692. https://doi.org/10.1371/journal.pone.0128692.

Lobell, David B., George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray. 2020. "Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis." *American Journal of Agricultural Economics* 102 (1): 202–19. https://doi.org/10.1093/ajae/aaz051.

Lu, Xin, Linus Bengtsson, and Petter Holme. 2012. "Predictability of Population Displacement after the 2010 Haiti Earthquake." *PNAS, Proceedings of the National Academy of Sciences of the United States of America* 109 (29): 11576–81. https://doi.org/10.1073/pnas.1203882109.

Maas, Paige, Shankar Iyer, Andreas Gros, Wonhee Park, Laura McGorman, Chaya Nayak, and P. Alex Dow. 2019. "Facebook Disaster Maps: Aggregate Insights for Crisis Response and Recovery." In *Conference Proceedings: 16th International Conference on Information Systems for Crisis Response and Management*, edited by Zeno Franco, José J. González, and José H. Canós, 836–47. Valencia, Spain: Polytechnic University of Valencia.

Marshall, Sarah. 2012. "Citizen Journalists Report Sierra Leone Elections by SMS." *Journalism*, November 20, 2012. https://www.journalism.co.uk/news/citizen-journalists-report-sierra-leone-elections-by-sms-/s2/a551240/.

Masaki, Takaaki, David Locke Newhouse, Ani Rudra Silwal, Adane Bedada, and Ryan Engstrom. 2020. "Small Area Estimation of Non-Monetary Poverty with Geospatial Data." Policy Research Working Paper 9383, World Bank, Washington, DC.

McCall, Becky. 2020. "COVID-19 and Artificial Intelligence: Protecting Health-Care Workers and Curbing the Spread." *Lancet Digital Health* 2 (4): e166–e167. https://doi.org/10.1016/S2589-7500(20)30054-6.

McGough, Sarah F., John S. Brownstein, Jared B. Hawkins, and Mauricio Santillana. 2017. "Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data." *PLoS Neglected Tropical Diseases* 11 (1): e0005295.

Mehrabi, Zia, Mollie J. McDowell, Vincent Ricciardi, Christian Levers, Juan Diego Martinez, Natascha Mehrabi, Hannah Wittman, et al. 2020. "The Global Divide in Data-Driven Farming." *Nature Sustainability* 4 (February 2021): 154–60. https://doi.org/10.1038/s41893-020-00631-0.

Milinovich, Gabriel J., Gail M. Williams, Archie C. A. Clements, and Wenbiao Hu. 2014. "Internet-Based Surveillance Systems for Monitoring Emerging Infectious Diseases." *Lancet Infectious Diseases* 14 (2): 160–68. https://doi.org/10.1016/S1473-3099(13)70244-5.

Milusheva, Sveta. 2020. "Managing the Spread of Disease with Mobile Phone Data." *Journal of Development Economics* 147 (November): 102559. https://doi.org/10.1016/j.jdeveco.2020.102559.

Milusheva, Sveta, Robert Marty, Guadalupe Bedoya, Elizabeth Resor, Sarah Williams, and Arianna Legovini. 2020. "Can Crowdsourcing Create the Missing Crash Data?" In *COMPASS '20: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 305–06. New York: Association for Computing Machinery. https://doi.org/10.1145/3378393.3402264.

Oliver, Nuria, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé, et al. 2020. "Mobile Phone Data for Informing Public Health Actions across the COVID-19 Pandemic Life Cycle." *Science Advances* 6 (23): eabc0764. https://doi.org/10.1126/sciadv.abc0764.

PAHO (Pan American Health Organization) and WHO (World Health Organization). 2016. "Zika Cases and Congenital Syndrome Associated with Zika Virus Reported by Countries and Territories in the Americas: Cumulative Cases, 2015–2016." PAHO, Washington, DC. https://www.paho.org/hq/dmdocuments/2016/2016-dec-29-phe-ZIKV-cases.pdf.

Peak, Corey M., Amy Wesolowski, Elisabeth zu Erbach-Schoenberg, Andrew J. Tatem, Erik Wetter, Xin Lu, Daniel Power, et al. 2018. "Population Mobility Reductions Associated with Travel Restrictions during the Ebola Epidemic in Sierra Leone: Use of Mobile Phone Data." *International Journal of Epidemiology* 47 (5): 1562–70.

Pepe, Emanuele, Paolo Bajardi, Laetitia Gauvin, Filippo Privitera, Brennan Lake, Ciro Cattuto, and Michele Tizzoni. 2020. "COVID-19 Outbreak Response: A Dataset to Assess Mobility Changes in Italy Following National Lockdown." *Scientific Data* 7: 230. https://doi.org/10.1038/s41597-020-00575-2.

Perisic, Igor. 2018. "How Artificial Intelligence Is Already Impacting Today's Jobs." *Economic Graph* (blog), September 17, 2018. https://economicgraph.linkedin.com/blog/how-artificial-intelligence-is-already-impacting-todays-jobs.

Perol, Thibaut, Michaël Gharbi, and Marine Denolle. 2018. "Convolutional Neural Network for Earthquake Detection and Location." *Science Advances* 4 (2): e1700578. https://doi.org/10.1126/sciadv.1700578.

Pratihast, Arun Kumar, Ben DeVries, Valerio Avitabile, Sytze De Bruin, Lammert Kooistra, Mesfin Tekle, and Martin Herold. 2014. "Combining Satellite Data and Community-Based Observations for Forest Monitoring." *Forests* 5 (10): 2464–89. https://doi.org/10.3390/f5102464.

Resch, Bernd, Florian Usländer, and Clemens Havas. 2018. "Combining Machine-Learning Topic Models and Spatiotemporal Analysis of Social Media Data for Disaster Footprint and Damage Assessment." *Cartography and Geographic Information Science* 45 (4): 362–76.

Reynard, Darcy, and Manish Shirgaokar. 2019. "Harnessing the Power of Machine Learning: Can Twitter Data Be Useful in Guiding Resource Allocation Decisions during a Natural Disaster?" *Transportation Research Part D: Transport and Environment* 77 (December): 449–63.

Ritchie, Hannah, and Max Roser. 2019. "Natural Disasters." *Our World in Data*. Global Change Data Lab and Oxford Martin Program on Global Development, University of Oxford, Oxford, UK. https://ourworldindata.org/natural-disasters.

Robinson, Bella Fay, Robert Power, and Mark Cameron. 2013. "A Sensitive Twitter Earthquake Detector." In *WWW '13: Proceedings of the 22nd International Conference on World Wide Web*, 999–1002. New York: Association for Computing Machinery. https://dl.acm.org/doi/10.1145/2487788.2488101.

Salathé, Marcel, Linus Bengtsson, Todd J. Bodnar, Devon D. Brewer, John S. Brownstein, Caroline Buckee, Ellsworth M. Campbell, et al. 2012. "Digital Epidemiology." *PLoS Computational Biology* 8 (7): e1002616.

Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, et al. 2019. "Aequitas: A Bias and Fairness Audit Toolkit." Cornell University, Ithaca, NY. https://arxiv.org/abs/1811.05577.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Serajuddin, Umar, Hiroki Uematsu, Christina Wieser, Nobuo Yoshida, and Andrew L. Dabalen. 2015. "Data Deprivation: Another Deprivation to End." Policy Research Working Paper 7252, World Bank, Washington, DC.

Servick, Kelly. 2020a. "Cellphone Tracking Could Help Stem the Spread of Coronavirus: Is Privacy the Price?" *Science*, March 22. https://www.sciencemag.org/news/2020/03/cellphone-tracking-could-help-stem-spread-coronavirus-privacy-price.

Servick, Kelly. 2020b. "COVID-19 Contact Tracing Apps Are Coming to a Phone Near You: How Will We Know Whether They Work?" *Science*, May 21. https://www.sciencemag.org/news/2020/05/countries-around-world-are-rolling-out-contact-tracing-apps-contain-coronavirus-how.

Stephens-Davidowitz, Seth. 2017. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are*. New York: HarperCollins.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." Proceedings of 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019.

Sylvester, Gerard, ed. 2019. "E-Agriculture in Action: Big Data for Agriculture." Food and Agriculture Organization of the United Nations and International Telecommunication Union, Bangkok. http://www.fao.org/3/ca5427en/ca5427en.pdf.

Tatem, Andrew J., Youliang Qiu, David L. Smith, Oliver Sabot, Abdullah S. Ali, and Bruno Moonen. 2009. "The Use of Mobile Phone Data for the Estimation of the Travel Patterns and Imported Plasmodium Falciparum Rates among Zanzibar Residents." *Malaria Journal* 8 (December): 287. https://doi.org/10.1186/1475-2875-8-287.

Tiecke, Tobias G., and Andreas Gros. 2016. "Connecting the World with Better Maps." *Facebook Engineering* (blog), February 22, 2016. https://engineering.fb.com/core-data/connecting-the-world-with-better-maps/.

Vale, Steven, and Taeke Gjaltema. 2020. "High-Level Group for the Modernisation of Official Statistics." United Nations Economic Commission for Europe, Geneva. https://statswiki.unece.org/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Official+Statistics.

Wallace, Eric, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. "Universal Adversarial Triggers for Attacking and Analyzing NLP." Cornell University, Ithaca, NY. http://arxiv.org/abs/1908.07125.

Wesolowski, Amy, Nathan Eagle, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. 2012. "Heterogeneous Mobile Phone Ownership and Usage Patterns in Kenya." *PLoS ONE* 7 (4): e35319. https://doi.org/10.1371/journal.pone.0035319.

WHO (World Health Organization). 2008. "The Top 10 Causes of Death." *Fact Sheets* (blog), May 24, 2008. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Williams, Kehinde, Adebayo Peter Idowu, and Emmanuel Olonade. 2015. "Online Road Traffic Accident Monitoring System for Nigeria." *Transactions on Networks and Communications* 3 (1): 10–30. https://doi.org/10.14738/tnc.31.589.

Wilson, Robin, Elisabeth zu Erbach-Schoenberg, Maximilian Albert, Daniel Power, Simon Tudge, Miguel Gonzalez, Sam Guthrie, et al. 2016. "Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake." *PLoS Currents* 8 (February 24). https://doi.org/10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c.

WRI (World Resources Institute). 2019. "Palm Oil Industry to Jointly Develop Radar Monitoring Technology to Detect Deforestation." Press release, October 31, 2019. https://www.wri.org/news/2019/10/release-palm-oil-industry-jointly-develop-radar-monitoring-technology-detect.

Yang, Shihao, Samuel C. Kou, Fred Lu, John S. Brownstein, Nicholas Brooke, and Mauricio Santillana. 2017. "Advances in Using Internet Searches to Track Dengue." *PLoS Computational Biology* 13 (7): e1005607.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David B. Lobell, Stefano Ermon, et al. 2020. "Using Publicly Available Satellite Imagery and Deep Learning to Understand Economic Well-Being in Africa." *Nature Communications* 11 (1): 2583. https://doi.org/10.1038/s41467-020-16185-w.

Zindi. 2020. "GIZ AI4D Africa Language Challenge, Round 2: $6,000 USD." *Competitions*, June 1, 2020. https://zindi.africa/competitions/ai4d-african-language-dataset-challenge.

Zou, James, and Londa Schiebinger. 2018. "AI Can Be Sexist and Racist: It's Time to Make It Fair." *Nature* 559 (7714): 324–26. https://doi.org/10.1038/d41586-018-05707-8.