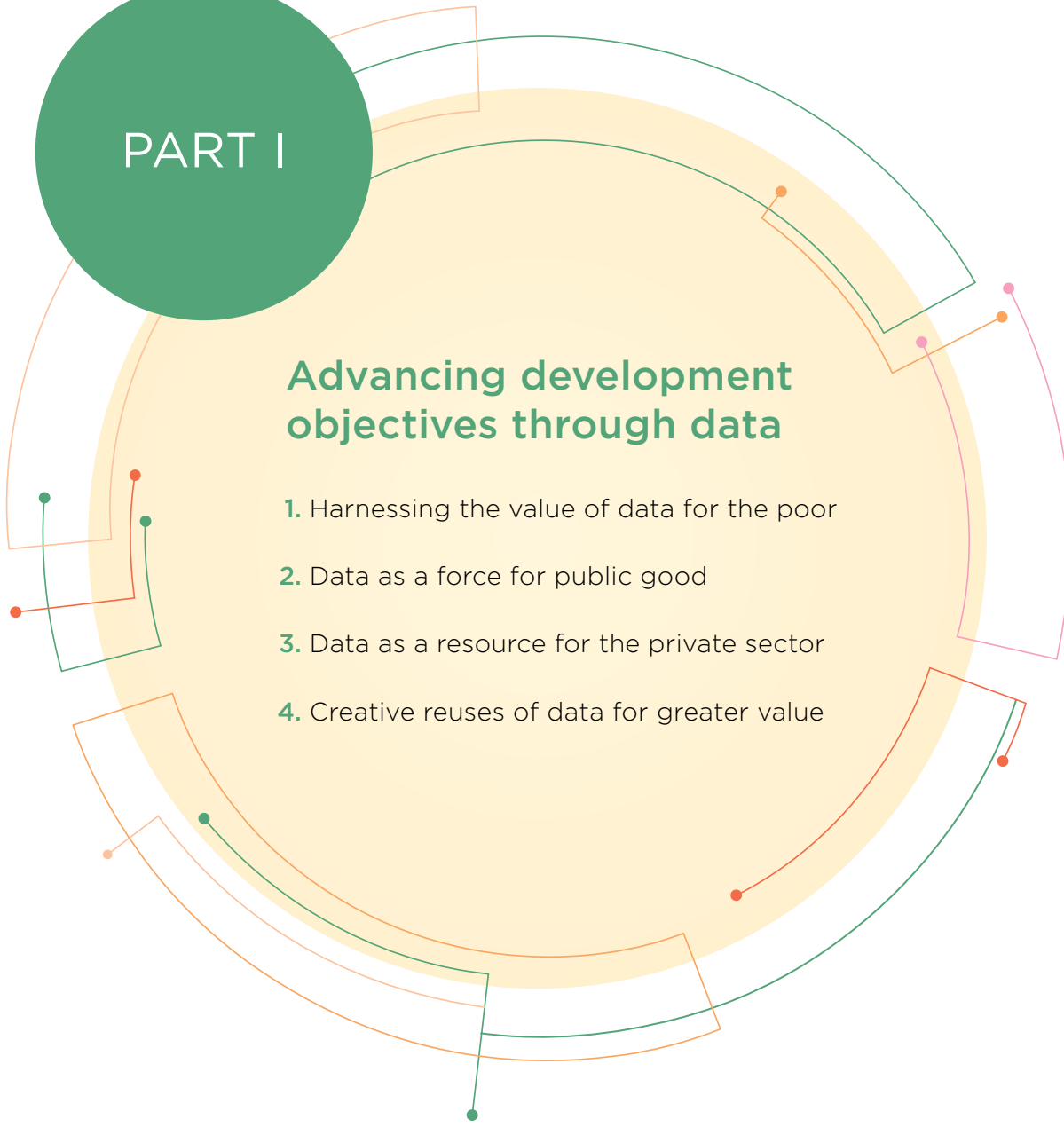




PART I



Advancing development objectives through data

1. Harnessing the value of data for the poor
2. Data as a force for public good
3. Data as a resource for the private sector
4. Creative reuses of data for greater value



Harnessing the value of data for the poor

Main messages

- 1 Data can improve people's lives in many ways. However, economic and political factors typically prevent benefits from being shared equitably.
- 2 The value of data for development is largely untapped. Realizing data's full value entails repeatedly reusing and repurposing data in creative ways to promote economic and social development.
- 3 The challenge is to develop a trust environment that safeguards against harmful misuse of data as they are exchanged between parties and enables data to be created, reused, and repurposed.
- 4 A strong data governance framework, composed of appropriate policies, laws, regulations, and institutions, is needed to ensure that the full value of data is realized and shared safely and equitably.



The untapped potential of data to serve development objectives

At the turn of the nineteenth century, English sociologist Seebohm Rowntree interviewed a sample of families with the aim of better understanding the poverty experienced not only by those he interviewed, but also by everyone in the town of York.¹ The findings from this work changed pre-conceptions by revealing that poverty was pervasive

outside of London and by demonstrating that people cycled in and out of poverty over the course of their lives.

How to turn data into information and information into insights that can help the poor is at the heart of this Report (see box 1.1 on the use here of the term *data*). In the twenty-first century, data possess the power to be truly life-changing. Most of the new and fascinating ways in which data affect the lives of many of us worldwide are linked to people being able

Box 1.1 What this Report means by *data*

The term *data* is difficult to define. It has meant different things at different times, and in different disciplines. Originally simply defined as facts, the term slowly came to mean facts as they related to mathematical representations. Despite the changing nature of data, most people would not have thought of things such as pictures, sounds, or words as data even as recently as a few decades ago. But times have changed, and major advances in computing power, together with innovative thinking, have resulted in, for example, radiomics, the science of converting medical images into data that, once structured and analyzed, can help improve a patient's diagnosis and prognosis.^a Similarly, sound can now be digitized and analyzed to, for example, explore and better understand the galaxies.^b And the growing field of text analytics converts words (such as keywords from Google searches) into structured data that help us better understand many social phenomena.^c Because the evolving definition of data stems simply from technological advances in computing and creative thinking, it is challenging to provide a specific description of data that would not soon seem archaic or anachronistic.

In very general terms, Carrière-Swallow and Haksar point out that “data can be quantitative or qualitative in nature, and may be stored on analog (that is, paper, stone tablets) or digital media.”^d This view conforms with how this Report uses the term. Indeed, some data are still collected on paper in many countries. Processing these data—digitizing them and entering them in a spreadsheet or database—allows them to be more easily analyzed, but a digital format is not necessarily an attribute of data.

The Organisation for Economic Co-operation and Development (OECD) states broadly that data are “characteristics or information, usually numerical, that are collected through observation.” More specifically, data are “the physical representation of information in

a manner suitable for communication, interpretation, or processing by human beings or by automatic means.”^e Although this description aligns fairly well with how the term is used in this Report, a few distinctions are worth noting. Here, data are sometimes collected through observation, though they need not be. Data can be the result of digital transactions or simply by-products of our daily digital lives. Also, in this Report, data are not synonymous with information. Rather, data must be processed, structured, and analyzed to be converted into information. This semantic distinction between data and information emphasizes the critical role of improved data management, literacy, and analysis for extracting information, and creating value, from data.

An expansive description of data that resonates well with how the term is used in this Report is provided by the UK National Data Strategy:

When we refer to data, we mean information about people, things and systems. . . . Data about people can include personal data, such as basic contact details, records generated through interaction with services or the web, or information about their physical characteristics (biometrics)—and it can also extend to population-level data, such as demographics. Data can also be about systems and infrastructure, such as administrative records about businesses and public services. Data is increasingly used to describe location, such as geospatial reference details, and the environment we live in, such as data about biodiversity or the weather. It can also refer to the information generated by the burgeoning web of sensors that make up the Internet of Things.^f

a. Gillies, Kinahan, and Hricak (2015); Yala et al. (2021).

b. See, for example, Leighton and Petculescu (2016).

c. See, for example, Stephens-Davidowitz (2017).

d. Carrière-Swallow and Haksar (2019, 17).

e. Organisation for Economic Co-operation and Development (OECD), “Glossary of Statistical Terms: Data,” OECD Statistics Portal, <https://stats.oecd.org/glossary/detail.asp?ID=532>.

f. See “What We Mean by Data” (DCMS 2020).

to extract greater value from data. Indeed, the data produced by people can be used in innovative ways to help them, but one does not have to be the producer or user of data to benefit from the data revolution. In fact, the data often collected from a small sample of people can help shape policy to improve the lives of a vastly larger population, whether they were part of the sample or not—just as Seebohm Rowntree revealed in his pioneering efforts. But for such approaches to work, the samples must be truly representative of the population, including the poor and other marginalized groups. And yet both traditional censuses and sample surveys, as well as new data sources captured by the private sector, may fail to fully cover the most disadvantaged groups.

An important attribute of data is that using them does not diminish their value to be reused for some other purpose—data are inexhaustible. But reusing or repurposing data typically requires well-functioning data systems that facilitate the safe flow of data in formats that make the data valuable to many users. These systems, however, typically do not function well in many low- and middle-income countries.

Moreover, data have a dark side. Making data accessible to more users and creating systems that facilitate their reuse also opens the door for data to be misused in ways that can harm individuals or development objectives. With lives becoming increasingly intertwined in the digital world, each day brings new concerns about protecting personal data; misinformation; and attacks on software, networks, and data systems.

Well-functioning data systems thus balance the need to *safeguard* against outcomes that harm people, while simultaneously *enabling* the potential for data to improve lives. This Report returns often to the need to strike this balance between safeguarding and enabling.

The findings and recommendations in this Report are drawn from an extensive array of material, including academic research, international development agency reports, commercial experiences, and a series of consultations with innovators and stakeholders in the data world. Although this Report reinforces and builds on findings from *World Development Report 2016: Digital Dividends*,² the World Bank report *Information and Communications for Development 2018: Data-Driven Development*,³ and many reports on digital technology, this Report differs by focusing on how data themselves, rather than the adoption of digital technology, can improve the lives of poor people.⁴

World Development Reports often synthesize established findings from analytical work and research, but

the issues and content surrounding data are evolving rapidly. Many of the topics covered continue to be widely debated in rich and poor countries alike. Consensus has yet to emerge, and research is at an early stage, particularly on how these issues affect low- and middle-income countries. The goal, therefore, is not to be overly prescriptive, but to develop frameworks to help policy makers and countries think through the trade-offs and adopt a balanced approach to developing both safeguards and enablers. Countries should make the most of data, but safely, and as appropriate for their social, political, and economic context.

The growing literature on data over the last few years is largely written from a high-income country perspective.⁵ This Report therefore sets out to fill the large gap in the literature on the effects of data on poor people and poor countries.

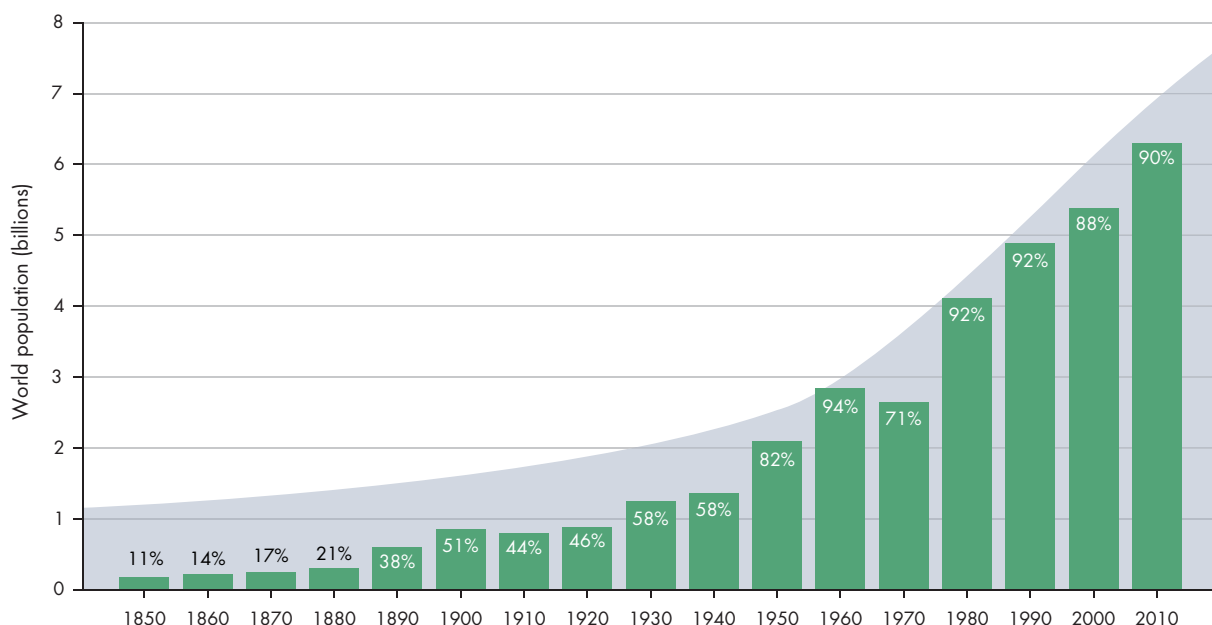
A brief history of data

Many of the themes of this Report were emerging even in the earliest days of data collection and use. For millennia, people have been collecting data. The oldest censuses date back to at least 2000–1000 BCE to ancient Egypt, Greece, and China, who enumerated people, livestock, and food items.⁶ The Romans fielded a census of men and their possessions every five years—a practice referenced in the Christian Bible.⁷

Over the long history of data collection, the type of data collected and the ways data have been used have changed as societies' priorities, values, power structures, and government objectives have changed.⁸ Record keepers in the Incan Empire between 1400 and 1500 CE counted people, dwellings, llamas, marriages, and potential army recruits.⁹ Rulers and administrators gave priority to counting sources of wealth and power considered of strategic importance (the data were kept secret from the public). They collected information first and foremost on property for taxation and men for military recruitment and labor force purposes, as well as enumerating newly conquered peoples and territories. With little reason to believe that the data being collected were meant to improve lives, distrust was widespread—it was not uncommon for citizens to resist being counted or having their possessions counted.¹⁰

The ascent of Enlightenment ideals in eighteenth-century Europe, with their emphasis on objective scientific inquiry, brought a shift in attitudes toward the role of data in society—from simply counting and registering phenomena to describing and understanding living conditions for society as a whole.¹¹ During this era, and under the influence of the leading

Figure 1.1 The share of people counted in a census grew from about 1 in 10 in 1850 to 9 in 10 today



Source: Whitby 2020. Data at http://bit.do/WDR2021-Fig-1_1.

Note: The shaded area represents the world's population; the bars indicate the percentage of the population that was enumerated each decade.

intellectuals of the day, notions of the rule of law (and accountability of states) evolved, a social contract between the individual and the state coalesced, and the Declaration of the Rights of Man and of the Citizen emerged.¹² These became the foundational elements of the current discussions of a social contract for data based on human rights (see chapter 6).

Starting in the late eighteenth century, governments of the emerging nation-states in Europe and North America established statistical agencies to publish official statistics on the state of the nation and to inform public discourse. European nations began systematically conducting full-fledged population censuses, and a decennial national census became a provision of the US Constitution. By the end of the nineteenth century, half of the world's population had been enumerated in censuses (figure 1.1).¹³

These advances also led to some of the innovations in statistics and social science research methods that enabled the rise of the sample survey. The earliest examples of sampling date back to the late seventeenth and early eighteenth centuries, but they lacked the theoretical foundations to justify the method.¹⁴ Sampling remained highly controversial throughout the nineteenth century, but methodological advances, especially the concept of random sample selection,

led to its gradual acceptance in the early twentieth century. A series of influential articles in the 1930s, 1940s, and 1950s filled the holes in the theoretical foundations of survey sampling around the same time that sampling frames with universal coverage became available.¹⁵ Sample surveys grew enormously popular, especially in the United States, quickly covering a wide range of topics.

Modern geospatial data systems developed along a similar timeline. Building on the much older science of cartography, this type of data is rooted in the thematic maps of the eighteenth and nineteenth centuries. Its goal was to relate geography to other types of information.¹⁶ A prominent early application was the spatial mapping of disease outbreaks—for example, of yellow fever in New York City at the end of the eighteenth century and especially of cholera in British and other European municipalities during the pandemics of the nineteenth century.¹⁷ Most prominent among those is the map of London by physician John Snow. During the 1854 cholera outbreak, Snow plotted cholera-related deaths in London together with the city's water pumps, identifying a high concentration of cases close to a pump on Broad Street and deducing that water from this pump was causing infections (map 1.1). New cases in the area stopped

almost entirely once the pump had been removed.¹⁸ Since the advent of Snow's map, innovations in printing and computer technology as well as the rise of remote sensing have made geospatial data and their applications versatile and ubiquitous.¹⁹

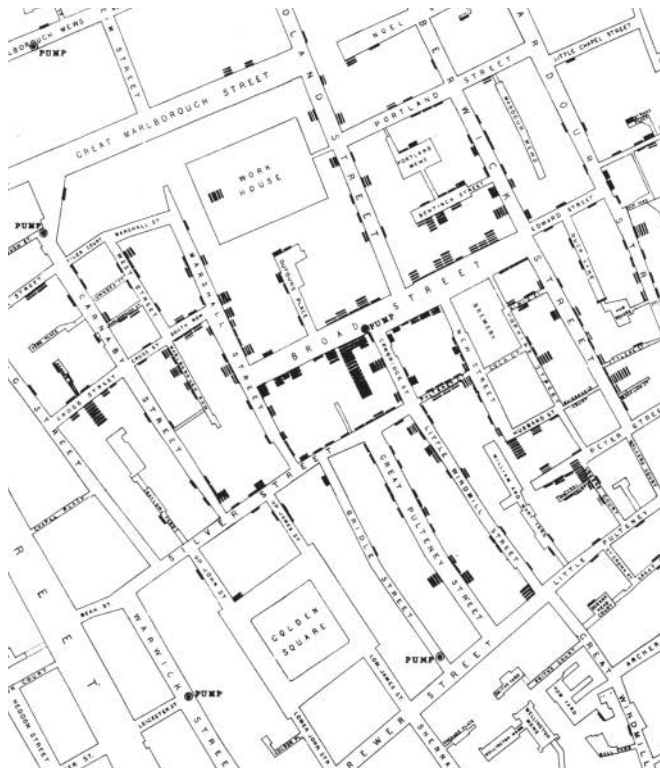
With the digital revolution, the types and scope of data have changed dramatically, and the volume of data collected has grown exponentially. In this new landscape, private sector actors are playing an increasingly larger role in data collection through platform-based business models in which data are collected passively as a by-product of business processes. Digital platforms have also expanded the opportunities for citizens to collect data, which often occurs when governments fail to collect data (see spotlight 1.1). Examples include Utunzi, a platform that allows individuals and organizations to report and document violence against LGBTQI individuals,²⁰ and various platforms that allow users to report air pollution levels, deforestation, and other location-specific environmental data to raise awareness and spur action.

The foundational origins of data protection laws can be linked to the Enlightenment era. Although there is a clear arc from these historical concepts of rights governing interactions between the state and the individual to principles guiding data protection, most policies guiding data regulation are very modern (see chapter 6). The principles of data protection can trace their immediate roots to the US Fair Information Practice Principles developed in the 1970s and that formed the basis for the 1980 OECD (Organisation for Economic Co-operation and Development) Guidelines (revised in 2013).²¹ Similarly, the basic substantive rights and obligations in the European Union's General Data Protection Regulation, reflected first in its 1995 Directive on the Protection of Personal Data, trace their roots to the OECD Guidelines.²²

A data typology

Although data can be used to improve development outcomes, the challenges differ across data types. To help readers conceptualize these data types and better understand those challenges, this Report sorts data types using a two-dimensional framework (table 1.1). In the first dimension, data are classified based on whether the original *intent* was for public or commercial purposes. Both new and traditional types of data collected for commercial purposes are called *private intent data*. Data originally collected for public purposes are called *public intent data*, regardless of the collection instrument or the entity that manages

Map 1.1 John Snow's innovative mapping of the cholera epidemic in London in 1854 revolutionized tracing of the disease



Source: Ball 2009. Map segment reproduced from John Snow, *On the Mode of Communication of Cholera*, 2nd ed. (London: John Churchill, 1855).

Note: The solid black rectangles of various sizes represent deaths from cholera.

the data. Public intent data tend to be collected purposefully with a view toward representativeness. By contrast, private intent data are self-selecting in that they cover only users of cellphones and the internet, for example, and collection of these data may be more incidental.

The second dimension distinguishes between “traditional” and “new” data collection methods. Public intent data are typically associated with traditional data types such as censuses and surveys, although newer sources of data (such as from satellite imaging or e-government platforms) have become more prevalent. By design, traditional data collection efforts by governments are for public purposes and are used to inform policy making. But because the collection of public data via traditional methods tends to be relatively costly,²³ surveys are conducted infrequently,²⁴ and they often lack the granularity necessary to make meaningful inferences about subpopulations of interest. Meanwhile, traditional public intent data offer important advantages over new private intent data



Table 1.1 Examples of data types based on original intent and collection methods

Data collection methods and tools	Public intent data	Private intent data
Traditional	Census, national accounts, household surveys, enterprise surveys, labor force surveys, surveys of personal finance, administrative records	Any survey conducted by private entities, including public opinion surveys deployed by private entities; administrative data from company financial accounts
New	Location data from satellite imaging, digital identification, facial recognition from public cameras, public procurement data from e-government platforms	Just-in-time digital data on individual behavior/choices from digital platforms in the private sector

Source: WDR 2021 team.

in terms of their coverage of the population—and thus their potential to benefit more people—and their format, which makes them amenable to inferential analytics by researchers and government officials.

Private intent data are often associated with new sources of data produced using digital tools and applications that are growing rapidly. Compared with traditional public intent data, new private data sources offer greatly improved timeliness, frequency, and granularity of data, but they may not be representative in coverage. New private intent data can contribute significantly to addressing public sector development challenges. Private intent data collected through cellphones, internet usage, satellites, remote sensors, and other sources provide information about individuals and geographic locations that traditional surveys simply cannot.

Any simple framework used to classify data types carries limitations. Although much public intent data have long been collected using traditional methods, those methods are being updated and adapted. The new methods will increasingly supplement or replace traditional methods, and so the traditional–new differentiation in table 1.1 is likely to evolve. The distinction between public and private stewardship of data also may not be a salient one in some cases. For example, citizen-generated data—data that people or their organizations produce to directly monitor, demand, or drive change on issues that affect them—can be produced through crowdsourcing mechanisms or citizen reporting initiatives, and such data are often organized and managed by civil society groups. The data may reside with a private entity, but they are clearly collected for public purposes.

Although data gathered through new methods for private purposes offer tremendous potential to improve timeliness and detail through massive sample size, they are not a panacea for the shortcomings

of public intent data collected using more traditional methods. For one thing, private firms have little incentive to curate their data for sharing, and thus these data are not readily amenable for public use. A potentially more difficult challenge is coverage. Data collected for public policy purposes are almost always designed to represent the relevant current population (such as individuals, firms, health facilities, students, or schools). However, survey designers face challenges in meeting the representativeness objective in terms of both coverage (such as underrepresentation of slum inhabitants, top earners, or informal enterprises) and timeliness (due to delays in data processing). By contrast, collectors of private intent data rarely need or have an interest in full population coverage; they focus much more on specific subgroups (such as consumers and suppliers). Thus, even though sample sizes can be massive and very timely, they can provide only partial reflections of the population. A study from the United Kingdom examined data from a variety of social media platforms and found that none was representative of the population, particularly underrepresenting the elderly, the less well educated, and lower-income people.²⁵

Public policies and programs need to be informed by data that represent the relevant population. For this reason, private intent data should not be viewed as a substitute for public intent data in understanding the scope of many development problems (box 1.2). That said, the joint use of public intent data collected using traditional methods and newer sources of private intent data offers interesting opportunities to reap significantly more value added than the isolated use of one kind of data or the other. A key theme of this Report is that *governments should take advantage of complementarities between new and traditional data to confront development challenges*. For example, because the majority of the world's poor live in rural areas and derive

Box 1.2 Innovation in traditional surveys: A COVID-19 example in Brazil

A prime example of the importance of traditional surveys and their potential for innovation comes from Brazil. In May 2020, it was one of the first countries to complete nationally representative surveys to produce data on the prevalence of COVID-19.^a Fieldworkers clad in personal protective equipment conducted a serology test on randomly selected household members. This test detects the presence of antibodies in the blood as a response to a specific infection, such as COVID-19—that is, it detects the body’s immune response to the infection caused by the virus rather than the virus itself. While waiting for the results of the test, the fieldworkers administered a brief questionnaire to collect sociodemographic data and asked the tested household member whether she or he was experiencing symptoms associated with COVID-19.^b Asking questions about symptoms enabled the research team to estimate rates of asymptomatic infection. Sociodemographic questions, especially those about work and travel outside the home, enabled the team to measure how much a household member adhered to social distancing guidelines.

The test results were conveyed to the household member before the fieldworkers left the dwelling, and information on positive tests was sent to health authorities to help them track the spread of the virus. In May, 25,025 interviews in 133 “sentinel cities” were completed in the baseline survey. Cities were chosen because of their primacy in the local region as hubs of commerce and services for surrounding urban and rural areas. The survey was conducted three more times, the most recent round in late August 2020. Multiple survey rounds enabled researchers and public health officials to track the spread of the virus over time by region.

At least two findings based on these serology tests and the interviews are striking. First, COVID-19 infections were far more prevalent than had been recorded. Overall seroprevalence—the share of the population that tested positive for the pathogen—for the 90 cities with a sample size of 200 or greater was 1.4 percent in the baseline survey. Extrapolating this figure to the full population of

these cities, who represent 25 percent of the country’s population, produced an estimate of 760,000 cases, compared with the 104,782 cases reported for those cities in official statistics as of May 13, 2020. In the fourth round of the survey in August, the seroprevalence rate had climbed to 3.8 percent.^c

Second, there was a remarkably wide regional variation in seroprevalence around the 1.4 percent national average, ranging from less than 1 percent in most cities in the South and Center-West regions to 25 percent in the city of Breves in the Amazon (North region). Eleven of the 15 cities with the highest seroprevalence were in the North. The six cities with highest seroprevalence were located along a 2,000-kilometer stretch of the Amazon River. Beyond geography, seroprevalence varied across ethnic groups and was highest among indigenous populations (3.7 percent in the baseline survey). Understanding the scope of the overall problem and identifying regions and populations with the most pressing needs would not have been possible without population-based surveys. These data also provided information on the effectiveness (or lack thereof) of approaches adopted to combat the spread of the disease.

Broad support for investigating something as important and urgent as the prevalence of COVID-19 might have been expected, and yet opposition sprang up in some quarters. For example, in some areas sample size was suppressed by the rapid spread of disinformation through social media that characterized the interviewers as “swindlers,” or even as part of a plot to spread the virus. In 27 cities, interviewers were arrested, and in eight cities the tests were destroyed by the local police force.^d Overall, however, the example illustrates the importance of population-based surveys (and public intent data in general) for understanding the scope and nature of disease spread.

a. Hallal, Hartwig, et al. (2020). Brazil is the only country in Latin America to complete a national survey.

b. Hallal, Horta, et al. (2020).

c. UFPEL (2020).

d. Hallal, Hartwig, et al. (2020).

their livelihoods from the land, measuring agricultural productivity is central to policies and programs to eliminate extreme poverty. Yet recent research has shown that agricultural productivity, specifically crop yield, is poorly measured with traditional survey

approaches that rely on farmer-reported information on crop production and land areas.²⁶ When sample surveys rely instead on objective measurement methods, the resulting data not only accurately capture crop yields at surveyed locations, but also can be used

to inform and develop remote sensing models that combine data from surveys and satellites to provide highly localized crop yield estimates across entire regions and countries beyond the locations in which sample surveys are conducted.²⁷

Both public intent and private intent data have advantages and disadvantages and pose distinct challenges in terms of reuse and exchange to achieve development objectives. But because public intent and private intent data have inherent complementarities, they can be used jointly to bolster development. A ministry of health would be able to issue better public policy if it could connect its health data with that of other ministries such as education, labor, and planning, as well as with that of health providers, whether public or private, around the country. A private firm would be able to operate more effectively if it could link its data with other sources of information, such as satellite data on population density and socioeconomic data on wealth and well-being.

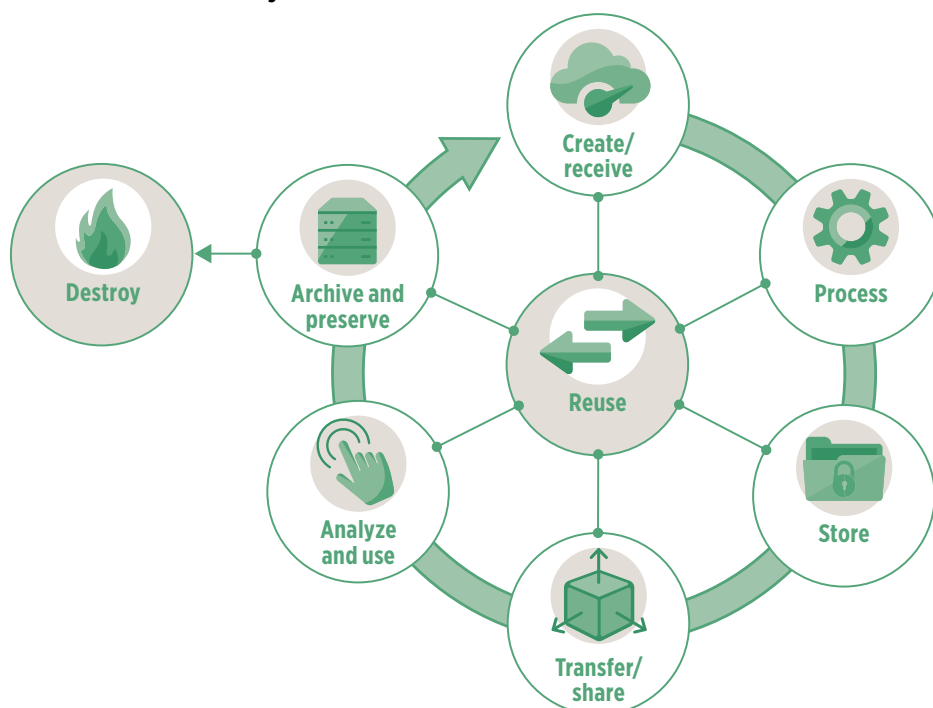
Combining the two types of data could advance evidence-based policy through more precise and timely official statistics that are produced more cheaply, while preserving the representativeness characteristic of public intent data. For example, building on the well-established infrastructure for socioeconomic surveys conducted by governments,

satellite data and call detail records from mobile phones offer new opportunities for updating poverty estimates for small areas more frequently. More generally, the high frequency of data collected for commercial purposes holds promise for producing better estimates of current socioeconomic conditions when large-scale, costly surveys such as censuses or integrated household surveys such as those of the World Bank's Living Standards Measurement Study are infrequent. Real-time data on prices, nighttime lights, or trade flows could be used to help “nowcast” (that is, generate an estimate for the current time based on data collected with a lag in time) macroeconomic data to avoid lags in availability.

The economics of data and political economy issues

The potential to extract further value from the proliferation of data is significant because data are inexhaustible or “nonrival”—that is, a person's call detail records, location history, internet usage, and medical records, among other things, can be used repeatedly by firms and governments for different purposes without depleting them.²⁸ This finding is illustrated by the data life cycle (figure 1.2), which depicts the potential circularity of data use, reuse, and

Figure 1.2 The data life cycle



Source: WDR 2021 team.

repurposing, as long as data can be made safely accessible across a wide array of users and unless explicit steps are taken to destroy the data.

Because of the nonrival character of data and the implications for their limitless reuse, it is inherently difficult to place an economic value on data, although many attempts have been made to do so. The diverse approaches taken range from a cost-based methodology that adds up different components of the information value chain;²⁹ to approaches that directly quantify the economic benefits that data yield by improving efficiency, reducing transaction costs, or expanding markets;³⁰ to estimates based on the stock market value of data-intensive companies and related acquisition transactions.³¹ Although all of these approaches agree on the high value of data, the partial nature of these estimates, together with the heterogeneity of the approaches, prevent any definitive conclusion.

In view of today's increasingly sophisticated application of machine learning and artificial intelligence to drive data-based innovations, it is quite conceivable that the economic value of unanticipated secondary uses of data may far exceed the value of the primary use of data—that is, the use for which they were originally collected. These characteristics raise the prospect of serious underinvestment in data collection from a social perspective because the entities bearing the cost of data collection will not necessarily be the ones capturing its full economic value.

At the same time, data are not a pure public good because they are excludable, allowing the entity that originally collects and holds the data to prevent others from accessing them. Examples abound across the public sector of the unwillingness of data holders to share data with other government entities and the public. In the private sector, firms may not want to sell or exchange their data with others, perhaps because governments and firms lack the capacity to share or exchange their data in a safe manner, or more simply because of a lack of incentives (or legal requirements) to make the data available. In some cases, reuse and sharing of data may cede informational advantages to competing firms in the private sector or rival entities in the public sector. Although the excludability of data suggests that they could readily be traded on markets, other economic characteristics complicate this classical approach to addressing allocation issues (see box 1.3).

A fundamental reason for the lack of incentives to share, sell, or exchange data is the considerable economies of scale that accrue to holding data and the associated economic or political power that they bring to

the data holder. Although the returns to the first few bits of data are essentially zero, there is a point past which the returns from additional data, and from improvements in the systems supporting these data, are substantial and increasing until they ultimately level off.³² For example, in the field of artificial intelligence the size of datasets is a critical determinant of the accuracy of predictive algorithms.³³ Modern deep-learning techniques, with their complex models, have an even more voracious appetite for large datasets than traditional machine learning, and they may not begin to experience diminishing returns until they incorporate much larger scales with datasets containing tens or hundreds of millions of data points.

In addition to economies of scale, data are characterized by economies of scope because combining different types of related datasets can yield insights that otherwise would have been unavailable from one type of data alone. Similarly, weak and seemingly very tangential relations can be identified through machine learning techniques with larger and larger volumes of data. For example, Google's search engine data may be used to evaluate the effectiveness of advertising on YouTube, which is also owned by Google.³⁴ Social media can also track users' behavior to then build very detailed advertising profiles.

Imbalances in information sharing, concentrations of power, and equity concerns: A dark side of data

Such strong economies of scale and scope in data, and the resulting assemblage of valuable information by some actors at the potential expense of those who are excluded from the transaction, may lead to a concentration of power—economic or political—in the hands of those with privileged access to large volumes of data.³⁵ In the private sector, market forces are likely to lead to data agglomeration and market concentration in data-driven businesses, which may preclude entry by small firms and eventually create conditions for the abuse of market power. Today, the firms that control the greatest agglomerations of data are among the world's largest. The concentration of personal information in a handful of companies raises concerns about market power and discrimination. A key theme throughout this Report is balancing the gains in efficiency that new data bring with such equity concerns. On the government side, data agglomeration may lead to a concentration of personal information, which can be used to amass and maintain political power, discourage dissent, and even discriminate against some population segments. Measures that limit and

Box 1.3 The challenges of trading data through markets

From an economic perspective, it seems plausible that access to data is best solved by first defining clear economic property rights over data and then allowing parties to trade in data. However, the limited nascent economic literature on this subject suggests that for two reasons these propositions are not as straightforward as they may initially seem.

First, legal and economic challenges confound the definition of property rights over data. A central issue is the ambiguity involved in allocating property rights between the data subject and the data collector, each of which has some legitimate claim to be the “data owner.” Present legal frameworks such as the European Union’s General Data Protection Regulation allocate certain specific rights to the data subject, implicitly leaving residual rights to the data collector as a purely *de facto* property right.^a Typically, a greater degree of data protection will benefit the data owner to the detriment of other potential data users and vice versa. This finding suggests that there is an economically optimal level of data protection. However, without efficient allocation of property rights, this social welfare-maximizing outcome will not be attained.^b

The large synergies and complementarities that arise across different types of data (economies of scope) raise the concern that fragmented ownership patterns will prevent them from being realized, whether through strategic behavior or through technical barriers such as lack of interoperability. However, the classic trade-off between the static objective of making data widely available to maximize economic value and the dynamic objective of preserving incentives for further data to be collected^c has weakened considerably with the advent

of digital data that are often collected without cost as a by-product of other economic activities.

Second, although private bilateral market exchanges of data are well established in certain niches (specifically, trading personal data to target advertising), there are as of today no open multilateral markets for data, and many attempts to create such data markets have failed.^d Because data are one of many experience goods that are difficult to evaluate in advance in areas such as price and quality, an important challenge is how data providers can convey information about the quality of their data before providing access.^e

In practice, data provenance has become the main means of signaling the quality and accuracy of data, relying on the reputation of the original source. However, the metadata needed to establish provenance may themselves be subject to legal restrictions in areas such as privacy, and data sellers may have strategic incentives to conceal or manipulate such information. The theoretical literature demonstrates that the institutional mechanisms currently available for trade in data have led to a sharp trade-off between the feasible scale of a data market and the ability to verify the quality of the data traded.^f Data may be traded via markets on a much larger scale in the future, but legal and institutional adaptations will be crucial to address challenges regarding data property rights and quality.

a. Duch-Brown, Martens, and Mueller-Langer (2017).

b. Duch-Brown, Martens, and Mueller-Langer (2017).

c. Duch-Brown, Martens, and Mueller-Langer (2017).

d. Koutroumpis, Leiponen, and Thomas (2020).

e. This is known as the Arrow Information Paradox (Arrow 1962).

f. Koutroumpis, Leiponen, and Thomas (2020).

neutralize this kind of dominance founded on the control of data need to be central to any data governance framework.

Because reliable statistics can expose poor policy decisions and performance, dilute power, and increase public scrutiny and pressure on governments, vested interests can be expected to intervene to distort decisions about the collection, reuse, and sharing of data. And indeed this Report finds strong associations among country statistical performance, independence of national statistical offices, and freedom of the press, controlling for country size and income level (chapter 2). The patterns indicate that a free and empowered press is a critical check

on government power and an important facilitator of statistical independence and data transparency.

Alternative data sources can provide a check on political influences when the accuracy or impartiality of official statistics is in question. For example, online prices obtained through web scraping have been used to construct daily price indexes in multiple countries, providing a comparison with official inflation figures. Researchers found that from 2007 to 2011, when Argentina reported an average annual inflation rate of 8 percent, online data indicated that the rate exceeded 20 percent.³⁶ The higher figure was consistent with inflation expectations from household surveys conducted at the time and similar to estimates of

some provincial governments and local economists. Because online price data were available outside the country, efforts by Argentina's government to discourage local economists from collecting these data independently were largely ineffective. These practices were halted in 2015 as Argentina took steps to reaffirm its commitment to the transparency and reliability of official data through its National Institute of Statistics and Censuses (INDEC). Similar disparities between official inflation statistics and those obtained from online prices have recently emerged in Turkey.³⁷

Governments can pose broader challenges to the use of nonofficial data sources. For example, Tanzania's 2018 amendment to its 2015 Statistics Act threatened members of civil society groups that published independent statistical information with imprisonment. Approval of the National Bureau of Statistics was required to publish such information, and publishing statistics that "invalidate, distort or discredit" official statistics was deemed a criminal offense. These provisions were subsequently amended amid international pressure.³⁸

Finally, the transparency and reliability of official statistics can have important macroeconomic implications. At a time when public debt levels are exploding from pandemic-related spending (see spotlight 1.2), governments may be less than forthcoming with data on the public debt, potentially enabling them to overborrow and hide debts from both citizens and creditors, at least for a while. Eventually, however, that strategy can have negative repercussions. For example, in Mozambique three state-backed companies took on in 2013 and 2014 more than US\$2 billion in government-guaranteed debt, equivalent to about 13 percent of the gross domestic product (GDP).³⁹ Roughly US\$1.2 billion of it was borrowed without being disclosed to parliament and the public. The country's access to international credit markets was severely curtailed after the hidden loans were revealed in 2016. To rehabilitate its reputation, the government has undertaken a complex reform package to foster greater transparency and improve governance and anticorruption frameworks.⁴⁰

Data for development: A conceptual framework

This Report poses two fundamental questions. How can data better advance development objectives? And what kind of data governance arrangements are needed to support the generation and use of data in a safe, ethical, and secure way while also delivering value equitably? The first part of this Report identifies

the multiple pathways through which data can support or inhibit the development process, relying on the conceptual framework presented in this chapter, together with concrete illustrations and examples from recent experience in less developed and emerging countries.

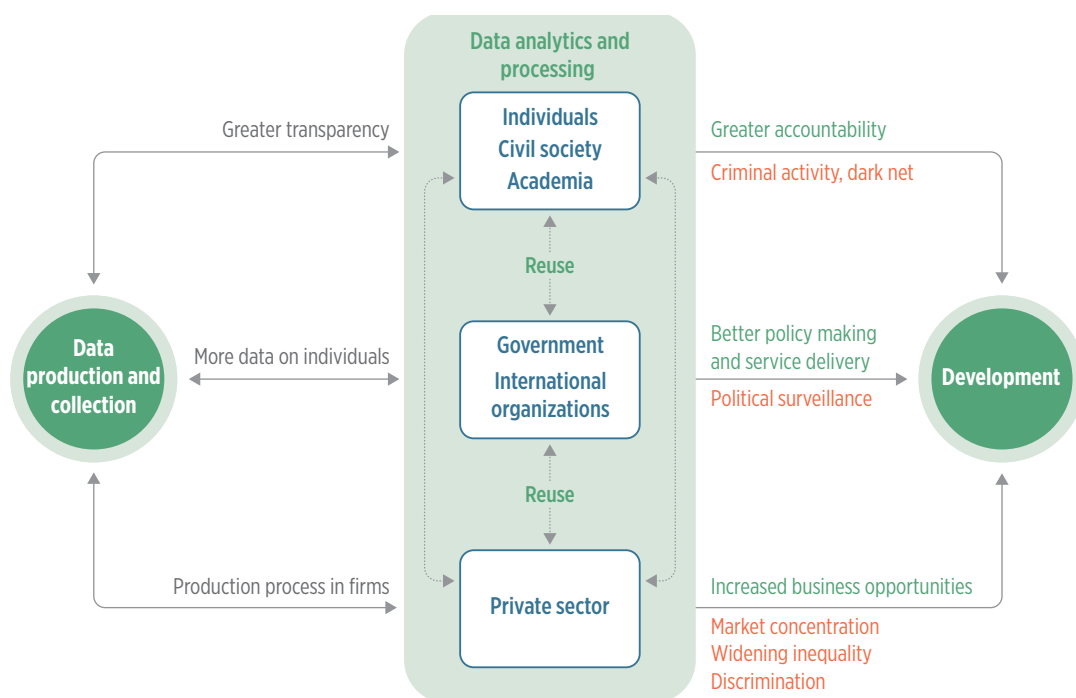
Three pathways by which data can support development

Data can contribute to development by improving the lives of the poor through multiple pathways. The conceptual framework that guides this Report focuses on three such horizontal pathways (figure 1.3). The middle pathway is data generated by or received by governments and international organizations to support program administration, service delivery, and evidence-based policy making (see chapter 2). The top pathway is data created and used by civil society and academia to monitor and analyze the effects of government programs and policies and by individuals to empower and enable them to access public and commercial services tailored to their needs. The bottom pathway is data generated by private firms. These data can be a factor of production that fuels firm and economic growth. But data also can be part of production processes in other ways (as an intermediate input, an output, or a by-product) and can be mobilized and repurposed to support development objectives (see chapters 3 and 4).

In figure 1.3, two-way arrows link data production and collection with the three groups of actors in the center of the figure. These arrows indicate that data do not merely flow to the actors. They also must be collected with purpose, and data processing and analytics by those actors provide important feedback about what data should be produced and collected going forward. The rectangle that encapsulates the actors indicates their centrality in processing and analyzing data to provide insights that lead to better lives and better development outcomes. Among these three pathways, data can be shared and exchanged flowing vertically across public, private, and civil society channels for further impact on development. Data collected for use in one of these pathways can be accessed and repurposed for a different use through other pathways or by other data users.

Government and international organizations. At a basic level, data enable governments to understand the impact of policies and improve program administration and service delivery. For traditional data types such as household and firm surveys, national accounts, and administrative data, governments (or agents authorized by governments) have been central to collection efforts. They have collected data typically

Figure 1.3 Three pathways along which data can foster development



Source: WDR 2021 team.

Note: Positive impacts are shown in green; negative impacts are shown in red.

for specific purposes, often intended to improve policies and encourage development. However, without strong data systems in place to support data analysis in relevant applications, much of the potential for data to improve outcomes is unrealized.

Important factors in supporting successful national data systems include trained staff, budgetary autonomy for agencies that collect data, adequate data infrastructure, connected databases, and international partnerships (see chapter 9). However, these resources are often scarce in low-income countries, leaving these countries the least equipped to collect and effectively use the data necessary to assess and understand the scope and nature of the development problems they face and make inroads to solving them. Enhancing the capacity of client countries to collect, analyze, and utilize data therefore has been, and will continue to be, a priority of the World Bank Group, and it is a major focus of this Report. International organizations can help countries to address lack of funding, technical capacity, governance, and demand for public intent data and to overcome these barriers. Sovereign-supported multilateral and bilateral development institutions are also important collectors and disseminators of data in their own right, and they support country governments in their efforts to improve and deploy data better.

A better ability to exchange public intent data across many platforms (interoperability) could increase their impact on development. Despite their advantages in coverage, suitability for some types of analysis, and potential for informing and improving policy, public intent data are often stored in different government agencies and formatted in different ways. Fragmentation and incompatibilities thus limit a government's scope to use its data to the fullest extent to improve policies, service delivery, and targeting. Interoperability across public intent data sources is therefore an important goal.

The central role of government and international organizations in fostering development through data use and reuse is captured in figure 1.3 by the placement of this pathway in the center of the figure (see chapter 2).

Individuals, civil society, and academia. In the top pathway, making data widely available enables individuals and civil society to hold governments accountable for policy choices. Inputs from civil society provide a feedback mechanism through which policies can be adapted and improved, leading to more responsive governance. Civil society organizations themselves create data by collecting surveys and crowdsourcing information directly from citizens. Such data can foster discussion, government

accountability, and transparency. Simply providing individuals with better access to their own data collected by government, international, or private sector actors is another way to enable citizens to advocate for themselves and improve their lives.

This pathway includes the use of administrative datasets by academic researchers to improve the quantity and quality of available evidence on social programs and policies.⁴¹ For example, administrative linked employer-employee datasets have been used to document earnings inequality and to study the sources of its decline in Brazil⁴² and to study underreporting of wages by formal firms⁴³ and the effects of business start-up programs in Mexico.⁴⁴ Often carried out in partnership with firms or governments, this type of research is being published increasingly in top academic journals.⁴⁵ However, broadening researchers' access to administrative datasets remains a challenge, even in countries with well-developed statistical systems.⁴⁶

The private sector. Through the bottom pathway, data generated by the private sector also hold promise for improving the lives of the poor (see chapter 3). For one thing, data have become critically important in the production process of many firms. Indeed, the business models of some of the world's largest firms (such as Amazon, Google, and Facebook) are predicated on data. Some important platform business models emerging in middle-income countries (such as Grab in Indonesia and Mercado Libre in Latin America) could greatly expand market access opportunities for small and medium enterprises. Other data-based private solutions can directly improve the lives of poor people—such as digital credit, often applied for via cellphone, which facilitates financial inclusion. Private financial services providers are also using alternative credit scoring techniques that take advantage of users' digital footprints to train machine learning algorithms to identify, score, and underwrite credit for individuals who otherwise would lack documentation of their creditworthiness.

Data reuse, sharing, and repurposing for all pathways. Enabling data reuse and repurposing is central to realizing their value (see chapter 4). Such reuse can take place between actors within each of the three pathways, but also across pathways. The two-way arrow in figure 1.3 between private firms and government indicates the reuse and repurposing for public policy of data originally collected for commercial purposes and the reuse and repurposing of public intent data by firms. Similarly, the two-way arrow between individuals/civil society/academia and governments indicates the reuse, sharing, and repurposing of data between

those parties. The final two-way arrows reflect the use of private sector data and data-driven applications by individuals/civil society/academia and the use of data and analysis generated by individuals/civil society/academia by firms.

The many examples of repurposing data to improve development outcomes include using geo-spatial location data from mobile phones, mobile call detail records, or social media (Facebook) and online search (Google) data to predict and trace the outbreak of disease, especially COVID-19 (box 1.4).⁴⁷ Online media and user-generated content can be used to map water/flood events in real time for water management and food security. Combining satellite imagery data from private and public sources can be used to monitor crop yields and forecast malnutrition.⁴⁸

The COVID-19 experience has also shown how public statistics constructed from private sector data—on credit card spending, employment, and business revenues—can serve as a new tool for empirical research and policy analysis. In the United States, indicators disaggregated by ZIP code, industry, income group, and business size showed that small businesses and low-income workers providing in-person services within wealthier ZIP codes were hardest-hit by the reduction in consumer spending during the crisis.⁴⁹ The patterns suggest that widespread tax cuts or relief checks are not effective when people are afraid to go out and spend. Unemployment insurance benefits and grants or low-cost loans targeting struggling businesses are likely a better approach.⁵⁰

Ways in which the same three pathways can harm development

Although use, reuse, and repurposing of data offer great prospects for fostering development, they simultaneously pose significant risks that must be managed to avoid negative development impacts. The mounting nature of such concerns has prompted calls for a new social contract around data. These risks can manifest themselves through public, private, and civil society pathways. Thus figure 1.3 also presents some concrete (though by no means exhaustive) illustrations (in red) of such negative impacts through each of the three pathways.

In the middle pathway, governments can abuse citizens' data for political ends. As public sector data systems improve and become increasingly interoperable, governments may accumulate a wide array of information about specific individuals. As long as public accountability is strong and state actors can be presumed to act in the broader public interest, this need not be a major concern. However, if those

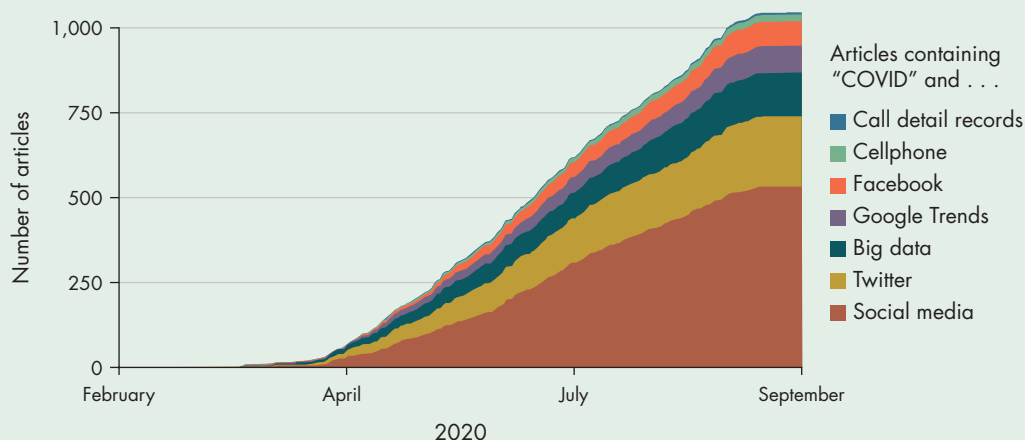
Box 1.4 Using private intent data to tackle COVID-19

At 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, proved extremely valuable in quantifying the effectiveness of policies ranging from partial curfews to strict lockdowns.^a These data enabled 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 inform epidemiological modeling. 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, travel restrictions and lockdowns on mobility had different effects on mobility across socioeconomic groups. Those in the top decile of wealth reduced their mobility up to twice as much as those in the bottom decile.^c

Despite the potential of deploying mobility data in the fight against COVID-19, their impact on policy has been limited in lower-income countries. Bottlenecks include a lack of technical expertise among government organizations; restrictions on data access, especially from mobile network operators; and a lack of investment and political will required to scale up onetime projects.^d

A review of the academic literature produces a broader look at the impact of repurposed data on the study of COVID-19 (figure B1.4.1 and map B1.4.1). 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 private intent data to track the disease and to offer policy and operational solutions (figure B1.4.1). Despite the relatively large number of articles in a short time frame, the coverage of lower-income countries was quite limited, especially in Africa (map B1.4.1). This pattern holds after adjusting

Figure B1.4.1 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-B1_4_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.

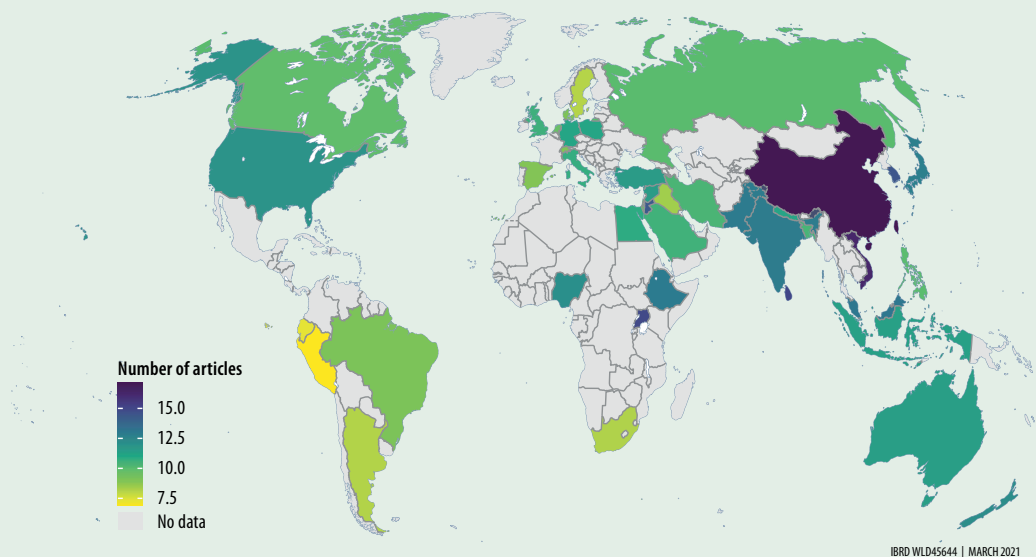
(Box continues next page)

presumptions do not hold, significant perils arise. One clear risk is the potential to misuse such data for politically motivated surveillance or discrimination along the lines of ethnicity, religion, race, gender, disability status, or sexual orientation. Another concern is the possible use of data by political incumbents,

domestic political players, or even foreign actors to unduly influence electoral processes by privately targeting misinformation to marginal voters during campaigns. Civil society actors can also misuse data for surveillance (to recruit members for violent extremism, for example) or to unduly affect electoral

Box 1.4 Using private intent data to tackle COVID-19 *(continued)*

Map B1.4.1 Use 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-B1_4_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.

the number of articles for death rates associated with COVID-19 in each country, and it likely reflects the difficulties in accessing data and the limited research funding and capacity.

- a. Oliver et al. (2020).
- b. Klein et al. (2020).
- c. Fraiberger et al. (2020).
- d. Oliver et al. (2020).

processes, as can private firms, but governments are more likely to do so. Again, these examples of misuse of data are mentioned to be indicative rather than exhaustive of all possibilities.

In the top pathway, individuals and organized groups can inflict considerable harm through cybercriminals who steal and manipulate sensitive information. The so-called dark net is a vast parallel network of hidden websites that can only be accessed using specific software tools and private authorizations. It acts as an underground digital platform for a wide array of criminal activities, facilitating illegal trade in drugs, counterfeit currency, stolen goods, credit card numbers, forged papers, firearms, and human organs. In addition to facilitating criminal activity in the real world, the internet can be the locus for cybercrime in the digital world, with burgeoning

security breaches leading to the theft of critical data and raising the risk of major disruptions to critical services. One recent study estimated the annual cost of such crime as between US\$57 billion and US\$110 billion in the United States alone.⁵¹ Data service providers have a tendency to underinvest in cybersecurity because the economic consequences of any data security breach are largely borne by the clients whose data are compromised.

In the bottom pathway, private firms can potentially abuse consumers' data through anticompetitive practices. Data-driven platform businesses experience steeply increasing returns to scale as user communities expand, leading to positive network externalities that make them more and more attractive to additional users. This dynamic has led to strong market concentration in platform businesses—including



e-commerce, search engines, and social media—raising concerns about abuse of market power. For services that are provided free of charge, abuse of dominance may manifest itself in declining quality of service, particularly in terms of the level of privacy offered to consumers. In other cases, use of algorithms can facilitate price collusion (tacit or otherwise). More broadly, data-driven businesses may exploit their vast information about consumer preferences and behavior to engage in aggressive or manipulative marketing techniques based on microtargeting of persuasive messages—a practice known as nudging—which may unduly influence consumers’ choices or simply be a nuisance.⁵²

Just as data can be reused for positive purposes, collecting and sharing sensitive data for ill-intentioned purposes can pose significant risks. For example, researchers at Cambridge Analytica developed a technique to map personality traits based on what people had “liked” on Facebook. The researchers paid users small sums to take a personality quiz and download an app that would scrape some private information from their profiles and those of their friends—an activity permitted at the time. Cambridge Analytica eventually obtained files for roughly 30 million users that contained enough information for the company to match users to other records and build psychographic profiles. However, only about 270,000 users—those who participated in the quiz—had consented to having their data harvested.⁵³ The outcome was that political campaigns were able to microtarget their political ads to individuals based on these profiles.

Although social media data can be reused to affect election outcomes, it is challenging to do so, and there is little solid evidence that the approach has had such effects thus far.⁵⁴ However, the Cambridge Analytica example demonstrates how private sector data can be leveraged by third parties (in this case, a political party) to attempt to influence voting behavior in ways the originators of the data (Facebook users and their friends) never intended.

The Cambridge Analytica example also highlights the importance of transparency as data are increasingly created, used, reused, and repurposed by a wider range of people, organizations, businesses, and other parties. At the most basic level, documentation of sources and collection and aggregation methods are crucial for data quality and for inspiring trust among users of data. But transparent documentation is not a priority in all countries, and some governments may consciously opt for data opacity, thereby significantly undermining public trust. In short, data policy

options are fraught with complex political economic constraints.

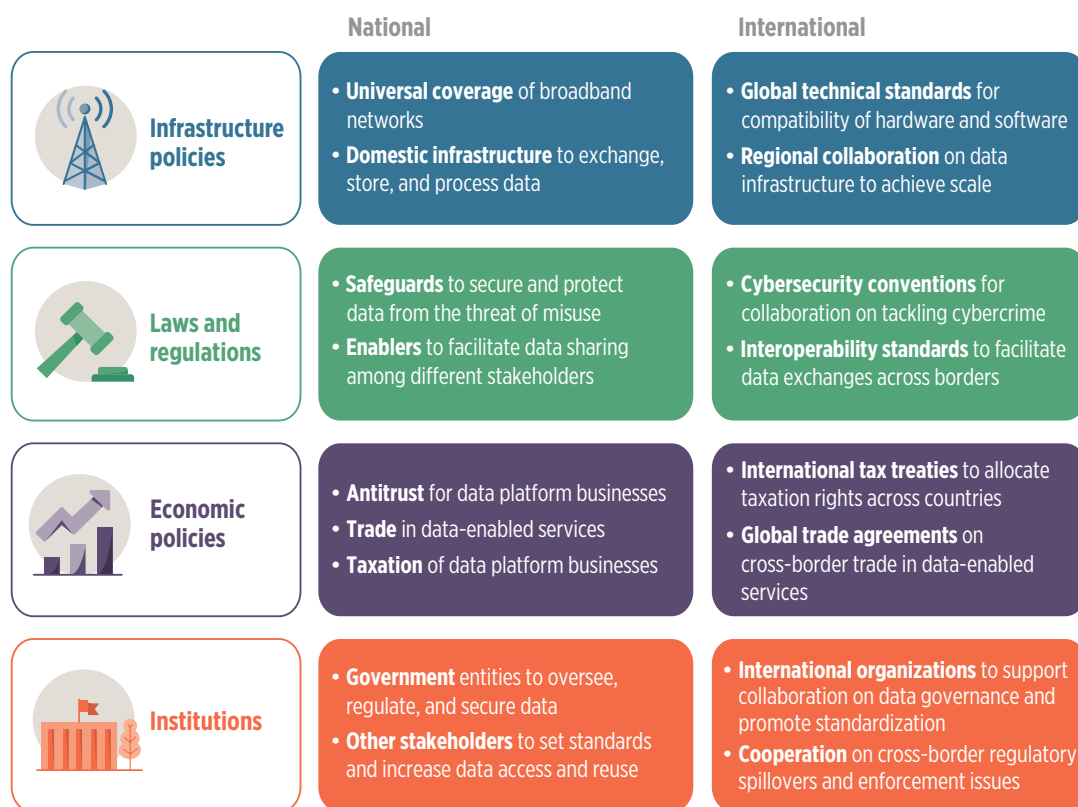
A data governance framework to enforce the social contract for data

Data governance entails creating an environment of implementing norms, infrastructure policies and technical mechanisms, laws and regulations for data, related economic policies, and institutions that can effectively enable the safe, trustworthy use of public intent and private intent data to achieve development outcomes. By providing predictability and confidence that these rights are being protected and protections are enforceable, a robust and effectively implemented data governance framework can strengthen *trust* in the data system, thereby incentivizing the use of data-driven products and services, increasing their *value*, and ensuring a more *equitable* distribution of benefits. In effect, data governance enforces the social contract around data, by applying the principles of trust, value, and equity.

A data governance framework can be visualized as four distinct layers that build on and support one another (figure 1.4). The foundational layer is the policy framework for data infrastructure—both the policies that promote universal access to internet data services and the policies that ensure that countries have adequate infrastructure to exchange, store, and process data efficiently over the internet. The next layer consists of the legal and regulatory environment for data itself, which creates rules to enable the reuse and sharing of data while safeguarding against their potential abuse and misuse. This normative framework for data interacts in significant ways with wider economic policy issues represented in the third layer, which affect a country’s ability to harness the economic value of data through competition, trade, and taxation. The fourth layer is the institutional ecosystem that ensures that data can deliver on their potential and that laws, regulations, and policies are effectively enforced.

Infrastructure policies. The digital character of modern data makes infrastructure indispensable for collecting, exchanging, storing, processing, and distributing data (see chapter 5). Individual access to data infrastructure is a prerequisite for both contributing one’s own data and accessing the data of others. The powerful network benefits, which arise as more and more people are connected to data infrastructure, are the economic underpinning for universal service policies that have also been widely endorsed

Figure 1.4 Data governance layers at the national and international levels



Source: WDR 2021 team.

politically. Significant policy efforts are needed to ensure adequate coverage of last-mile infrastructure that keeps pace with constantly evolving technological standards of performance. Policy makers also need to consider how to address demand-side factors, such as the affordability of handheld devices and data charges, as well as people's limited data literacy skills.

At the country level, affordable processing of data transactions and adequate speed call for increasingly sophisticated data infrastructure. A starting point is to facilitate the creation of internet exchange points that allow internet service providers to exchange domestic internet traffic across their respective networks without incurring expense and slowing speed by routing traffic overseas. A further step is to create a policy environment suitable for investment in colocation data centers. Such centers allow storage and retrieval of vast volumes of digital data, including local replicas of popular global internet content, and they can be used to provide dedicated access to overseas cloud computing capacity that facilitates increasingly sophisticated data processing and machine learning

techniques. Small-scale, regulatory deficiencies and inadequate competition all conspire to hold back the development of all forms of data infrastructure in many low- and middle-income countries, posing particular challenges for policy makers.

Laws and regulations. Legal and regulatory frameworks for data need to be adequately developed, with a balanced emphasis on both safeguards and enablers (see chapter 6). The legal and regulatory provisions to safeguard personal and nonpersonal data differ greatly because these two types of data are typically generated, used, and treated in very different ways.

Personally identifiable data convey information that is specific to a known individual, although identifiers (such as names, addresses, and social security numbers) that directly or indirectly point to a person (or entity) could be deleted.⁵⁵ Some types of personal data, such as health histories or banking transactions, may be more sensitive than others, such as shopping records. Nonpersonal data are generated about non-human subjects, including institutions or machines. They may include data on prices, traffic patterns,



weather, and agricultural practices. In practice, the boundary between personal and nonpersonal data is becoming increasingly blurred as it becomes possible to infer personal characteristics from nonpersonal data, such as mobile phone records. Advances in artificial intelligence also are making the deidentification of personal data more challenging and making personal inferences from combining multiple sources of nonpersonal data possible, thereby blurring the boundaries between personal and nonpersonal data.

The nature of data safeguards for personal data versus nonpersonal data is quite different. For personal data, a rights-based approach to data protection is appropriate, emphasizing the rights of data subjects as well as the obligations of data users as the primary considerations. For nonpersonal data, intellectual property rights provide the relevant frame of reference, and there is greater scope to weigh the balance of economic interests between safeguards and enablers. Another important and underdeveloped aspect of data protection, affecting both personal and nonpersonal data, is cybersecurity.

Complementing such safeguards, greater access to data for reuse can be enabled by open data regulations and by provisions that ensure technical interoperability between different types of data, allowing them to be readily combined and repurposed. Data portability provisions, which allow individuals to move their own data from one service provider to another, also help enhance the agency of data subjects.

Economic policies. Because of the proliferation of data-driven platform business models, the design of legal frameworks for data governance carries significant implications for the real economy that are often overlooked (see chapter 7). Competition agencies grapple with the market power of globally dominant technology firms operating data-driven business models. Tax authorities struggle to collect revenues from platform businesses that often have scale in a market without any physical presence and can readily shift tax liabilities across international borders. Trade policy introduces tensions between the need to protect data domestically and the desire to benefit from a growing cross-border trade in data-based goods and services. In each case, decisions about the design of the domestic regulatory framework for data will materially affect economic performance.

Institutions. For effective enforcement of the normative framework, a suitable institutional ecosystem that encompasses both state and nonstate actors must be in place (see chapter 8). The proliferation of arrangements around the world suggests that there is no single institutional blueprint for the

implementation of data governance frameworks. The important thing is to identify the critical functions needed to deliver on the safeguards and enablers embedded in legal statutes. Depending on the country context, it may make sense to assign some of these roles to existing institutions (such as the national statistical office or relevant sector regulators) or to create new institutions (such as data protection agencies or data intermediaries). Whatever the institutional architecture, common challenges facing the effective implementation of data governance policies include capacity and resource constraints, lack of institutional autonomy, difficulties adopting a data-driven culture, and problems of coordination across stakeholder groups.

International dimension. Although they are rooted in the domestic environment, data governance frameworks also have important international dimensions (as shown in figure 1.4 and further detailed in spotlights 7.2 and 8.1). In many instances, international treaties provide the overarching legal framework for the development of domestic legislation and regulations. International agreements are also critical in reaching resolution of long-standing data policy challenges such as how to treat cross-border data flows in international trade or how to allocate taxation rights for data transactions. At the institutional level, decisions made by policy makers and regulators, particularly in the larger global markets, will have important spillover effects in smaller countries, particularly those with which the markets have strong economic ties. These effects underscore the importance of cross-border cooperation in addressing common data governance challenges such as the regulation of market power in data-driven businesses. At the same time, data infrastructure is to a considerable extent cross-border in nature, with large volumes of data flowing to overseas storage and processing facilities and opportunities for regional collaboration around infrastructure development. Facilitation of such cross-border data movements also entails global harmonization of technical standards.

In addition, there is an important role for international cooperation in creating common standards and guidelines for statistical activities (spotlight 2.2). The creation of international measurement standards and protocols helps improve comparability of measures across countries in a way that allows national policy makers to understand their country's performance relative to that of their neighbors. Cross-country measurement of progress toward policy goals and, more generally, of statistical performance ensure that countries can benchmark and monitor their

data achievements and identify and strengthen their weaknesses. Good data governance, both at the national and international levels, ensures that the various components work together to enable the effective and safe use of data in order to extract value in a trustworthy, equitable way.

Putting it all together: Establishing an integrated national data system

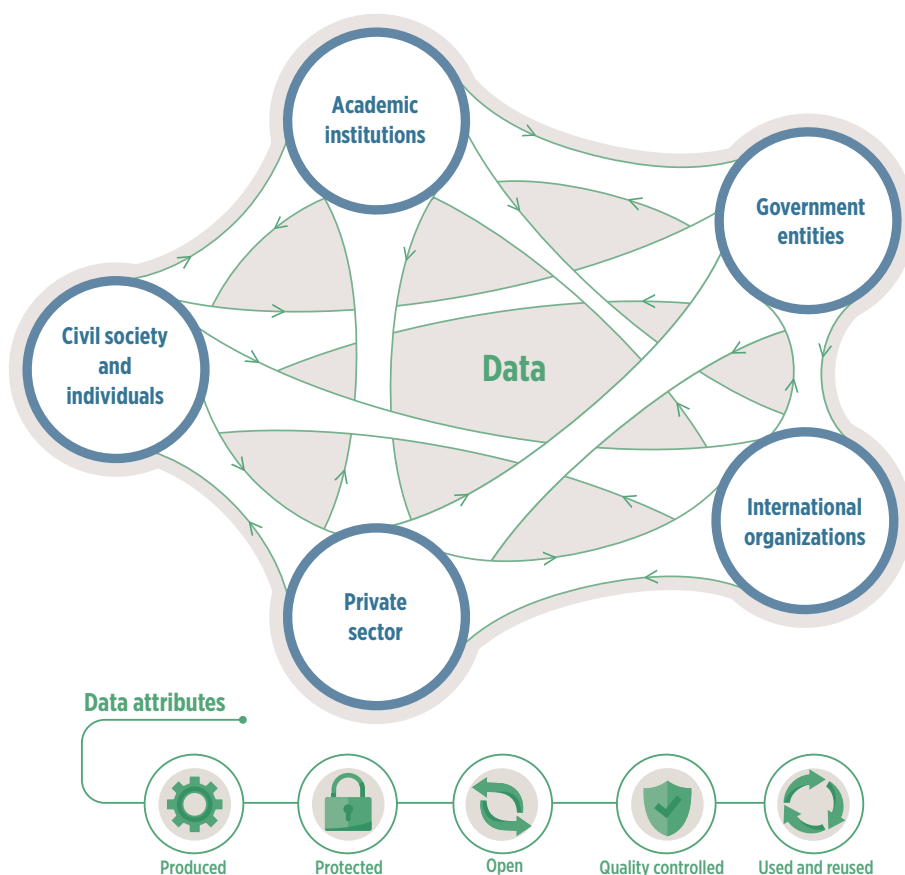
A well-functioning data governance framework ensures that infrastructure, laws and regulations, policies, and institutions work together to support the use of data in a way that aligns with the social contract for data. This framework defines the rules, and the associated compliance mechanisms, for how data can be safely used, reused, and shared by all stakeholders, including government entities, international organizations, civil society and individuals, academic

institutions, and the private sector. To realize data's potential, this framework must be built around a data system that not only ensures that data transactions are safe, but also actively promotes access to data by all stakeholders (figure 1.5).

This Report concludes with an aspirational vision of an integrated national data system (INDS) that can deliver on the promise of producing high-quality data and then making data open in a way that it is both protected and accessible to be shared and reused by all stakeholders (chapter 9). Such an aspirational INDS works seamlessly with the governance structure. If the governance system is viewed as establishing the rules of the road (and the institutions governing those rules), the INDS can be envisioned as a network of highways that connects all users and ensures the safe passage of data to and from destinations.

A well-functioning INDS is powered by people: people to produce, process, and manage high-quality data; people to populate the institutions that

Figure 1.5 Data flow safely across all stakeholders in an integrated national data system



Source: WDR 2021 team.



safeguard and protect the data against misuse; and people to draft, oversee, and implement data strategies, policies, and regulations. The system also needs all people to have sufficient skills and knowledge to use data in ways that allow them to hold the public and private sectors accountable. All this requires robust data literacy within data institutions, government ministries, the private sector, and the general population.

There is no singular blueprint for how to build an INDS. It certainly must be funded sufficiently to implement the infrastructure and institutions necessary for the system to function well. Incentives need to be in place to produce, protect, and share data, and to create a data culture in which people demand transparency and accountability. But how countries move toward this vision of an INDS will depend on their current capacity and the parameters of the social contract for data. Although the path toward an INDS will differ for each country, this Report proposes a sequenced maturity model to help assess progress and identify areas for more attention to further the development of a well-functioning INDS.

The maturity model is based on a progression of three stages: first, establishing fundamentals; second, initiating data flows; and third, optimizing the system (chapter 8). Although progress within these stages will differ by country—and for a given country progress in dealing with certain types of data also may differ—these three stages nonetheless serve as a useful reference to help assess weak spots and gaps in the construction of an INDS.

Establishing fundamentals first requires taking stock to identify the existing data types and the data processing activities carried out by different actors. This analysis should focus on activities already taking place—both inside and outside of government—that present potential development opportunities for data use, reuse, and repurposing, along with risks. Uncovering gaps in the stock of data or bottlenecks in gaining access to these data can help prioritize efforts to address gaps and remove barriers. Governments should also engage with the private sector and civil society stakeholders to develop legislation, rules, and standards to safeguard data, while encouraging data collection, processing, and use. Other steps in establishing fundamentals include efforts to facilitate public-private data sharing and cross-border data transfers by establishing contracts with information management services (such as identification systems) or licenses for regulated entities (such as banks and telecom operators) that create provisions for secure, protected data transactions between public

and private actors. Ensuring that the fundamentals are in place also includes developing a data governance strategy with policies and laws that promote the objectives of the INDS and enforce compliance with rules.

The next phase is to ensure that *data begin to flow across all the stakeholders*. One path to this goal is to establish a government agency with sufficient power to leverage compliance across ministries and public sector agencies in how they manage and exchange data. In addition, the rules and standards that enable greater interoperability among datasets must be established. Creating interoperability allows for innovative new uses of multiple data files as these data become accessible to a more diverse set of users. It also allows for the development of measurement standards to ensure data quality.⁵⁶ Public-private and cross-border data flows can be encouraged through multistakeholder engagements with domestic and international actors to promote harmonization principles, standards, and practices. Such engagements are particularly important for data protection and cybersecurity, which require coordination to be effective.

To reach the *optimized stage*, the tools and methods that helped create data flows should be incorporated into a unified whole-of-government approach. Ongoing, recurrent investments in training increase the effective use of data for decision-making and accountability. Similarly, recurrent investments in infrastructure keep systems sufficiently modern and expand access. Data quality, data integration, and data synchronization should be integral parts of all processes at this stage. Meanwhile, the safe flow of data through the data system should be continually assessed and stress tested for weakness.

Organization of this Report

This Report is divided into three parts. Part I identifies the multiple channels through which data can support or impede the development process, making sense of the data landscape and pointing out the associated development opportunities and risks. This part provides a conceptual framework (figure 1.3), together with illustrations and examples from recent experience in low- and middle-income countries.

Part II, which describes the data governance layers presented in figure 1.4, focuses on data governance broadly defined to include data infrastructure policy (chapter 5), the legal and regulatory framework for data (chapter 6), the related economic policy implications (chapter 7), and institutions (chapter 8). These diverse elements are effectively the building blocks

of a social contract that seeks to deliver the potential value of data equitably while safeguarding against harmful outcomes. Examples and case studies illustrate both the importance of establishing safeguards to prevent the misuse of data that could harm development objectives and how data can be better enabled to further development objectives.

Part III brings together the building blocks of the Report to present the vision of an integrated national data system (chapter 9).

Throughout the Report, spotlights at the end of chapters highlight relevant cases in low- and middle-income countries and internationally and explore various policy issues in more depth.

This Report was prepared against the backdrop of the COVID-19 pandemic. The pandemic itself is a vivid illustration of the usefulness of data in dealing with obstacles to development and the complexity of the associated governance challenges. Examples of how countries have used data as part of their response to COVID-19 are featured in chapters, using boxes and narratives to illustrate many of the issues addressed in the Report. Those issues include the deficiencies of public sector data systems and the complementarities between public intent and private intent data, as well as the legal and regulatory issues posed by accessing private intent data for public purposes. More broadly, through a discussion of the many ways in which data can help economic development, this Report aims to describe the challenges to realizing these gains, offer guidance on how to attain them, and propose safeguards for protecting citizens.

Notes

1. Rowntree (2000 [1901]).
2. World Bank (2016).
3. World Bank (2019).
4. The Report also builds on other themes featured in past World Bank reports, including the importance of building the data capacity of countries (see World Bank 2018). More generally, World Bank reports have long emphasized the importance of data, information, and knowledge for economic, social, and political development (see, for example, World Bank 2002). What has changed is the nature and amount of data available, the ways in which they are produced, and the ease with which they can be exchanged, reused, and shared to address development objectives. Thus the focus of this *World Development Report* is on data for better lives, particularly for the poor.
5. See, for example, OECD (2013, 2016, 2018a, 2018b, 2019).
6. Whitby (2020).
7. Grajales et al. (2013).
8. Thorvaldsen (2017).
9. Bethlehem (2009).
10. Thorvaldsen (2017).
11. de Heer, de Leeuw, and van der Zouwen (1999).
12. Conseil constitutionnel, “Déclaration des Droits de l’Homme et du Citoyen de 1789” [Declaration of Human and Civic Rights of 26 August 1789], Paris, <https://www.conseil-constitutionnel.fr/le-bloc-de-constitutionnalite/declaration-des-droits-de-l-homme-et-du-citoyen-de-1789>.
13. Whitby (2020).
14. Bethlehem (2009).
15. Bethlehem (2009).
16. Musa et al. (2013).
17. Wallis and Robinson (1987).
18. Musa et al. (2013).
19. Dempsey (2012).
20. LGBTQI stands for lesbian, gay, bisexual, transgender, queer (or questioning), intersex.
21. OECD (2013).
22. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (1995 Directive on Personal Data Protection, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31995L0046>) was repealed and replaced in 2016 by Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (EU GDPR, <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679>).
23. Kilic et al. (2017).
24. Serajuddin et al. (2015).
25. Blank and Lutz (2017).
26. Abay et al. (2019); Carletto, Jolliffe, and Banerjee (2015); Desiere and Jolliffe (2018); Gourlay, Kilic, and Lobell (2019).
27. Lobell, Azzari, et al. (2020); Lobell, Di Tommaso, et al. (2020).
28. Jones and Tonetti (2020). Treating data as a nonrival input in a production function draws on the earlier literature that modeled information and ideas as nonrival inputs to production. See Romer (1990) and Radner and Stiglitz (1984).
29. For example, Statistics Canada (2019) estimated the value of data in Canada considering the direct labor cost of data production, associated indirect labor costs, and other related expenses such as human resources management and financial control. It quantified the total own-account investment in databases in 2018 as between Can\$8 billion and Can\$12 billion.
30. Two examples illustrate this approach. First, the US Department of Commerce (2014) found that government data helped US businesses generate at least US\$24 billion a year. Second, Deloitte (2017) conducted a review of studies of the economic value of open data (public data available with no restrictions to users) to a wide range of users in the United Kingdom and found that satellite data from Landsat were worth US\$2 billion a year in commercial applications, while public transport routing and scheduling data from Transport for London

generated economic benefits to passengers valued at £80 million a year.

31. Several illustrations of this approach suggest the magnitude of monetary benefits. PwC (2019) found that market capitalizations of data-intensive companies are twice as likely to be in the top industry quartile as those of companies operating in the same sector that are not data-intensive. Li, Nirei, and Yamana (2019) looked at the sums paid for acquisitions of data-intensive firms and their impact on the market capitalization of the acquiring firm. Frier (2018) examined the revenue streams that companies are able to generate from sales of data or associated advertising revenues, finding that Apple charges application developers a commission of 30 percent of their sales for accessing its consumer data, which has earned the company US\$42.8 billion in sales over the past 10 years.
32. This is similar to suggesting that there is a nonconcavity in the value of data and information. It is also linked to the point that because ideas are nonrivalrous, they exhibit increasing marginal returns over a range. See Radner and Stiglitz (1984) and Romer (1990).
33. Juba and Le (2019).
34. Goldfarb and Tucker (2019).
35. Zingales (2017) notes that as the economic scale of firms becomes large in relation to governments, economic and political power may converge.
36. Cavallo (2013); Cavallo and Rigobon (2016).
37. Erkoyun (2020).
38. Nyeko (2019).
39. *Economist* (2019).
40. IMF (2019).
41. Cole et al. (2020).
42. Alvarez et al. (2018); Menezes-Filho et al. (2008).
43. Kumler, Verhoogen, and Frias (2020).
44. Kaplan, Piedra, and Seira (2011).
45. Chetty (2012); Cole et al. (2020).
46. Card et al. (2010).
47. Wesolowski et al. (2015).
48. Burke and Lobell (2017); Osgood-Zimmerman et al. (2018).
49. Chetty et al. (2020).
50. Rosalsky (2020).
51. CEA (2018).
52. Amnesty International (2019); Zuboff (2019).
53. Rosenberg, Confessore, and Cadwalladr (2018).
54. Hern (2018).
55. Kayaalp (2017).
56. Anyone wondering about the importance of establishing comparable definitions and developing precise instruments for these measures need only look at the US National Institute of Standards and Technology, established in 1901. It has been home to five Nobel laureates.

References

Abay, Kibrom A., Gashaw T. Abate, Christopher B. Barrett, and Tanguy Bernard. 2019. "Correlated Non-Classical Measurement Errors, 'Second Best' Policy Inference, and the Inverse Size-Productivity Relationship in

- Agriculture." *Journal of Development Economics* 139 (June): 171–84. <https://doi.org/10.1016/j.jdeveco.2019.03.008>.
- Alvarez, Jorge, Felipe Benguria, Niklas Engbom, and Christian Moser. 2018. "Firms and the Decline in Earnings Inequality in Brazil." *American Economic Journal: Macroeconomics* 10 (1): 149–89. <https://doi.org/10.1257/mac.20150355>.
- Amnesty International. 2019. "Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights." Report POL 30/1404/2019, Amnesty International, London. <https://www.amnesty.org/en/documents/document/?indexNumber=pol30%2f1404%2f2019&language=en>.
- Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, edited by National Bureau of Economic Research, 609–26. Princeton Legacy Library Series. Princeton, NJ: Princeton University Press.
- Ball, Laura. 2009. "Cholera and the Pump on Broad Street: The Life and Legacy of John Snow." *History Teacher* 43 (1): 105–19.
- Bethlehem, Jelke. 2009. "The Rise of Survey Sampling." Discussion Paper 09015, Statistics Netherlands, The Hague.
- Blank, Grant, and Christoph Lutz. 2017. "Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram." *American Behavioral Scientist* 61 (7): 741–56. <https://doi.org/10.1177/0002764217717559>.
- 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>.
- Card, David E., Raj Chetty, Martin S. Feldstein, and Emmanuel Saez. 2010. "Expanding Access to Administrative Data for Research in the United States." White Paper, National Science Foundation, Alexandria, VA. <http://www.rajchetty.com/chettyfiles/NSFdataaccess.pdf>.
- Carletto, Calogero, Dean Jolliffe, and Raka Banerjee. 2015. "From Tragedy to Renaissance: Improving Agricultural Data for Better Policies." *Journal of Development Studies* 51 (2): 133–48. <https://doi.org/10.1080/00220388.2014.968140>.
- Carrière-Swallow, Yan, and Vikram Haksar. 2019. "The Economics and Implications of Data: An Integrated Perspective." Departmental Paper 19/16, Strategy, Policy, and Review Department, International Monetary Fund, Washington, DC.
- Cavallo, Alberto. 2013. "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics* 60 (2): 152–65.
- Cavallo, Alberto, and Roberto Rigobon. 2016. "The Billion Prices Project: Using Online Prices for Inflation Measurement and Research." *Journal of Economic Perspectives* 30 (2): 151–78.
- CEA (Council of Economic Advisers). 2018. "The Cost of Malicious Cyber Activity to the U.S. Economy." CEA, White House, Washington, DC. <https://www.whitehouse.gov/wp-content/uploads/2018/02/The-Cost-of-Malicious-Cyber-Activity-to-the-U.S.-Economy.pdf>.

- Chetty, Raj. 2012. "Time Trends in the Use of Administrative Data for Empirical Research." Paper presented at NBER Summer Institute 2012, National Bureau of Economic Research, Cambridge, MA, July 2–27. http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf.
- Chetty, Raj, John N. Friedman, Nathaniel Hendren, Michael Stepner, and 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://www.nber.org/system/files/working_papers/w27431/w27431.pdf.
- Cole, Shawn, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. 2020. *Handbook on Using Administrative Data for Research and Evidence-Based Policy*. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab and Massachusetts Institute of Technology. <https://admindatahandbook.mit.edu/book/v1.0-rc6/index.html>.
- DCMS (Department for Digital, Culture, Media, and Sport, United Kingdom). 2020. "UK National Data Strategy." Policy paper, DCMS, London. <https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy>.
- de Heer, Wim, Edith Desirée de Leeuw, and Johannes van der Zouwen. 1999. "Methodological Issues in Survey Research: A Historical Review." *Bulletin of Sociological Methodology* 64 (1): 25–48.
- Deloitte. 2017. "Assessing the Value of TfL's Open Data and Digital Partnerships." Deloitte LLP, London. <http://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf>.
- Dempsey, Caitlin. 2012. "History of GIS." *GIS Lounge* (blog), May 14, 2012. <https://www.gislounge.com/history-of-gis/>.
- Desiere, Sam, and Dean Jolliffe. 2018. "Land Productivity and Plot Size: Is Measurement Error Driving the Inverse Relationship?" *Journal of Development Economics* 130 (January): 84–98. <https://doi.org/10.1016/j.jdevco.2017.10.002>.
- Duch-Brown, Nestor, Bertin Martens, and Frank Mueller-Langer. 2017. "The Economics of Ownership, Access, and Trade in Digital Data." JRC Digital Economy Working Paper 2017-01, Joint Research Center, European Commission, Seville, Spain. <https://ec.europa.eu/jrc/sites/jrcsh/files/jrc104756.pdf>.
- Economist. 2019. "The Net Tightens: A \$2bn Loan Scandal Sank Mozambique's Economy." August 22, 2019. <https://www.economist.com/middle-east-and-africa/2019/08/22/a-2bn-loan-scandal-sank-mozambiques-economy>.
- Erkoyun, Ezgi. 2020. "Researchers Say New Model Shows Turkish Inflation Well Above Official Tally." Reuters, October 22, 2020. <https://www.reuters.com/article/turkey-economy-inflation-int-idUSKBN2771EY>.
- Fraiberger, Samuel P., Pablo Astudillo, Lorenzo Candeago, Alex Chumet, 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." ArXiv:2006.15195 [Physics. soc-ph], July 27, Cornell University, Ithaca, NY.
- Frier, Sarah. 2018. "Is Apple Really Your Privacy Hero?" *Bloomberg Businessweek*, June 8, 2018. <https://www.bloomberg.com/news/articles/2018-08-08/is-apple-really-your-privacy-hero>.
- Gillies, Robert J., Paul E. Kinahan, and Hedvig Hricak. 2015. "Radiomics: Images Are More Than Pictures, They Are Data." *Radiology* 278 (2): 563–77. <https://doi.org/10.1148/radiol.201511169>.
- Goldfarb, Avi, and Catherine Tucker. 2019. "Digital Economics." *Journal of Economic Literature* 57 (1): 3–43. <https://doi.org/10.1257/jel.20171452>.
- 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://www.sciencedirect.com/science/article/pii/S0304387818306588>.
- Grajales, Carlos Gómez, Eileen Magnello, Robert Woods, and Julian Champkin. 2013. "Great Moments in Statistics." *Significance* 10 (6): 21–28.
- Hallal, Pedro Curi, Fernando P. Hartwig, Bernardo L. Horta, Gabriel D. Victora, Mariângela F. Silveira, Cláudio José Struchiner, Luis Paulo Vdaleti, et al. 2020. "Remarkable Variability in SARS-CoV-2 Antibodies across Brazilian Regions: Nationwide Serological Household Survey in 27 States." *medRxiv* (May 30). <https://www.medrxiv.org/content/10.1101/2020.05.30.20117531v1>.
- Hallal, Pedro Curi, Bernardo L. Horta, Aluísio J. D. Barros, Odir A. Dellagostin, Fernando P. Hartwig, Lúcia C. Pellanda, Cláudio José Struchiner, et al. 2020. "Trends in the Prevalence of COVID-19 Infection in Rio Grande do Sul, Brazil: Repeated Serological Surveys." *Ciência & Saúde Coletiva* 25 (supplement 1): 2395–401. <https://doi.org/10.1590/1413-81232020256.1.09632020>.
- Hern, Alex. 2018. "Cambridge Analytica: How Did It Turn Clicks into Votes?" *Guardian*, May 6, 2018. <https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>.
- IMF (International Monetary Fund). 2019. "Republic of Mozambique: Diagnostic Report on Transparency, Governance, and Corruption." IMF Country Report 19/276, IMF, Washington, DC. <https://www.imf.org/en/Publications/CR/Issues/2019/08/23/Republic-of-Mozambique-Diagnostic-Report-on-Transparency-Governance-and-Corruption-48613>.
- Jones, Charles I., and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review* 110 (9): 2819–58. <https://doi.org/10.1257/aer.20191330>.
- Juba, Brendan, and Hai S. Le. 2019. "Precision-Recall Versus Accuracy and the Role of Large Data Sets." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (01): 4039–48. <https://doi.org/10.1609/aaai.v33i01.33014039>.
- Kaplan, David S., Eduardo Piedra, and Enrique Seira. 2011. "Entry Regulation and Business Start-Ups: Evidence from Mexico." *Journal of Public Economics* 95 (11–12): 1501–15. <https://doi.org/10.1016/j.jpubeco.2011.03.007>.
- Kayaalp, Mehmet. 2017. "Modes of De-Identification." Paper presented at American Medical Informatics Association 2017 Annual Symposium, Washington, DC. November 6–8. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977668>.
- Kilic, Talip, Umar Serajuddin, Hiroki Uematsu, and Nobuo Yoshida. 2017. "Costing Household Surveys for Monitoring Progress toward Ending Extreme Poverty and



- Boosting Shared Prosperity.” Policy Research Working Paper 7951, World Bank, Washington, DC.
- 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>.
- Koutroumpis, Pantelis, Aija Leiponen, and Llewellyn D. W. Thomas. 2020. “Markets for Data.” *Industrial and Corporate Change* 29 (3): 645–60. <https://doi.org/10.1093/icc/dtaa002>.
- Kumler, Todd, Eric Verhoogen, and Judith Frías. 2020. “Enlisting Employees in Improving Payroll Tax Compliance: Evidence from Mexico.” *Review of Economics and Statistics* 102 (5): 881–96. https://doi.org/10.1162/rest_a_00907.
- Leighton, Timothy G., and Andi Petculescu. 2016. “Guest Editorial: Acoustic and Related Waves in Extraterrestrial Environments.” *Journal of the Acoustical Society of America* 140 (2): 1397–99. <https://doi.org/10.1121/1.4961539>.
- Li, Wendy C. Y., Makoto Nirei, and Kazufumi Yamana. 2019. “Value of Data: There’s No Such Thing as a Free Lunch in the Digital Economy.” RIETI Discussion Paper 19-E-022, Research Institute of Economy, Trade, and Industry, Tokyo. <https://www.rieti.go.jp/jp/publications/dp/19e022.pdf>.
- 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>.
- Lobell, David B., Stefania Di Tommaso, Calum You, Ismael Yacoubou Djima, Marshall Burke, and Talip Kilic. 2020. “Sight for Sorghums: Comparisons of Satellite- and Ground-Based Sorghum Yield Estimates in Mali.” *Remote Sensing* 12 (1): 100. <https://doi.org/10.3390/rs12010100>.
- Menezes-Filho, Naércio Aquino, Marc-Andreas Muendler, and Garey Ramey. 2008. “The Structure of Worker Compensation in Brazil, with a Comparison to France and the United States.” *Review of Economics and Statistics* 90 (2): 324–46.
- Musa, George J., Po-Huang Chiang, Tyler Sylk, Rachel Bavley, William Keating, Bereketab Lakew, Hui-Chen Tsou, and Christina W. Hoven. 2013. “Use of GIS Mapping as a Public Health Tool: From Cholera to Cancer.” *Health Services Insights* 6 (November): 111–16. <https://doi.org/10.4137/HSI.S10471>.
- Nyeko, Oryem. 2019. “Tanzania Drops Threat of Prison over Publishing Independent Statistics.” Human Rights Watch, Dispatches, July 3. <https://www.hrw.org/news/2019/07/03/tanzania-drops-threat-prison-over-publishing-independent-statistics>.
- OECD (Organisation for Economic Co-operation and Development). 2013. *The OECD Privacy Framework*. Paris: OECD. http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf.
- OECD (Organisation for Economic Co-operation and Development). 2016. “Big Data: Bringing Competition Policy to the Digital Era.” Report DAF/COMP(2016)14 (rev. November 29–30), Competition Committee, Directorate for Financial and Enterprise Affairs, OECD, Paris. [https://one.oecd.org/document/DAF/COMP\(2016\)14/en/pdf](https://one.oecd.org/document/DAF/COMP(2016)14/en/pdf).
- OECD (Organisation for Economic Co-operation and Development). 2018a. *Rethinking Antitrust Tools for Multi-Sided Platforms 2018*. Paris: OECD. <https://www.oecd.org/daf/competition/Rethinking-antitrust-tools-for-multi-sided-platforms-2018.pdf>.
- OECD (Organisation for Economic Co-operation and Development). 2018b. *Tax Challenges Arising from Digitalisation: Interim Report 2018*. Paris: OECD. <http://dx.doi.org/10.1787/9789264293083-en>.
- OECD (Organisation for Economic Co-operation and Development). 2019. *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies*. Paris: OECD. <https://www.oecd-ilibrary.org/content/publication/276aaca8-en>.
- 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>.
- Osgood-Zimmerman, Aaron, Anoushka I. Millea, Rebecca W. Stubbs, Chloe Shields, Brandon V. Pickering, Lucas Earl, Nicholas Graetz, et al. 2018. “Mapping Child Growth Failure in Africa between 2000 and 2015.” *Nature* 555 (7694): 41–47. <https://doi.org/10.1038/nature25760>.
- PwC. 2019. “Putting a Value on Data.” PwC, London. <https://www.pwc.co.uk/issues/data-analytics/insights/putting-value-on-data.html>.
- Radner, Roy, and Joseph E. Stiglitz. 1984. “A Nonconcavity in the Value of Information.” In *Bayesian Models in Economic Theory*, edited by Marcel Boyer and Richard E. Kihlstrom, 33–52. Studies in Bayesian Econometrics Series 5. Amsterdam: Elsevier.
- Romer, Paul M. 1990. “Endogenous Technological Change.” *Journal of Political Economy* 98 (5): S71–S102.
- Rosalsky, Greg. 2020. “The Dark Side of the Recovery Revealed in Big Data.” *Planet Money Newsletter*, October 27, 2020. <https://www.npr.org/sections/money/2020/10/27/927842540/the-dark-side-of-the-recovery-revealed-in-big-data>.
- Rosenberg, Matthew, Nicholas Confessore, and Carole Cadwalladr. 2018. “How Trump Consultants Exploited the Facebook Data of Millions.” *New York Times*, March 17, 2018. <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>.
- Rowntree, Benjamin Seebohm. 2000. *Poverty: A Study of Town Life*, 2d ed. Bristol, UK: Policy 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.
- Statistics Canada. 2019. “The Value of Data in Canada: Experimental Estimates.” *Daily*, July 10, 2019, Statistics Canada, Ottawa. <https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm>.

- Stephens-Davidowitz, Seth. 2017. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are*. Illus. ed. New York: Dey Street Books.
- Thorvaldsen, Gunnar. 2017. *Censuses and Census Takers: A Global History*. Routledge Studies in Modern History Series. London: Routledge. <https://doi.org/10.4324/9781315148502>.
- UFPEL (Federal University of Pelotas). 2020. "FAPESP e Todos pela Saúde garantirão a continuidade do estudo EPICOVID-19 BR." Coordenação de Comunicação Social, Pró-Reitoria de Gestão da Informação e Comunicação, UFPEL, Pelotas, Rio Grande do Sul, Brazil. <http://ccs2.ufpel.edu.br/wp/2020/08/31/fapesp-e-todos-pela-saude-garantirao-a-continuidade-do-estudo-epicovid-19-br/>.
- US Department of Commerce. 2014. "Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data." Office of the Chief Economist, Economics and Statistics Administration, US Department of Commerce, Washington, DC. <https://www.commerce.gov/files/fostering-innovation-creating-jobs-driving-better-decisions-value-government-data>.
- Wallis, Helen M., and Arthur Howard Robinson, eds. 1987. *Cartographical Innovations: An International Handbook of Mapping Terms to 1900*. Tring, UK: Map Collector Publications.
- Wesolowski, Amy, Taimur Qureshi, Maciej F. Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, et al. 2015. "Impact of Human Mobility on the Emergence of Dengue Epidemics in Pakistan." *Proceedings of the National Academy of Sciences* 112 (38): 11887–92.
- Whitby, Andrew. 2020. *The Sum of the People: How the Census Has Shaped Nations, from the Ancient World to the Modern Age*. New York: Basic Books.
- World Bank. 2002. *World Development Report 2002: Building Institutions for Markets*. Washington, DC: World Bank; New York: Oxford University Press.
- World Bank. 2016. *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank.
- World Bank. 2018. *Data for Development: An Evaluation of World Bank Support for Data and Statistical Capacity*. Washington, DC: Independent Evaluation Group, World Bank.
- World Bank. 2019. *IC4D, Information and Communications for Development 2018: Data-Driven Development*. Washington, DC: World Bank. <http://documents1.worldbank.org/curated/en/987471542742554246/pdf/128301-9781464813252.pdf>.
- Yala, Adam, Peter G. Mikhale, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, et al. 2021. "Toward Robust Mammography-Based Models for Breast Cancer Risk." *Science Translational Medicine* 13 (578): 1–11. <https://doi.org/10.1126/scitranslmed.aba4373>.
- Zingales, Luigi. 2017. "Towards a Political Theory of the Firm." *Journal of Economic Perspectives* 31 (3): 113–30.
- Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs. <https://www.hbs.edu/faculty/Pages/item.aspx?num=56791>.