Closing the Data Divide for a More Equitable U.S. Digital Economy

By Gillian Diebold | August 22, 2022

In the United States, access to many public and private services, including those in the financial, educational, and health-care sectors, are intricately linked to data. But adequate data is not collected equitably from all Americans, creating a new challenge: the data divide, in which not everyone has enough high-quality data collected about them or their communities and therefore cannot benefit from data-driven innovation. This report provides an overview of the data divide in the United States and offers recommendations for how policymakers can address these inequalities.

INTRODUCTION

Technological advances have made it cheaper and easier than ever to collect, process, and use data. This data helps individuals, businesses, and governments make better decisions, and data-driven innovation is a critical pathway for boosting social and economic prosperity. But, in a world in which economic opportunities, government services, and health-care outcomes are intricately linked to data, how individuals and communities are reflected in datasets and how they can use datasets about themselves significantly impact their ability to fully participate in the data economy. Divides are emerging between the data haves and the data have-nots, and these data divides can greatly impact individuals and communities. While many in academia, civil society, and the public sector have considered the impacts of the digital divide, such as disparities in access to broadband, mobile devices, or computers, few have explored the data divide or considered steps to address it.

The data divide refers to the social and economic inequalities that may result from a lack of collection or use of data about individuals or communities. Data divides can manifest in various ways. People in certain places may face greater environmental risks because an insufficient number of sensors gather data about their environmental conditions. Likewise, patients may receive inadequate medical treatments because their demographic is underrepresented in clinical trial data. Other times, some students receive suboptimal educational opportunities because school districts lack the systems to track and measure links between educational programs and outcomes. These data divides can emerge for different reasons, including a lack of resources, political pressure, or legal and regulatory issues.

As the Center for Data Innovation first wrote in its 2014 report, *The Rise of Data Poverty in America,* some Americans will be born in hospitals leveraging health informatics, attend schools powered by learning analytics, and live and work in "smart" communities that use data to maximize their economic, social, and environmental prosperity. But others won't, and the scarcity of data about themselves and their communities will mean that they will not benefit from the advantages of an increasingly data-driven world. These imbalances in data collection and use lead to data divides, and policymakers should prioritize addressing these data inequalities.

This report offers an overview of the data divide in the United States and recommends actions that policymakers should take to ensure the fair and equitable representation and use of data for all Americans. First, the report defines the data divide and contextualizes it within the conversation of the more-familiar digital divide. It identifies several areas where data divides persist, ranging from demographic and geographic data gaps to instances of inequitable data systems. Finally, it offers nine recommendations for how policymakers can address the data divide:

- Improve federal data quality by developing targeted outreach programs for underrepresented communities.
- Enhance data quality for non-government data.
- Ensure comparable data collection and monitoring methodologies among the government and civil society.
- Support increased utilization and incorporation of crowdsourced and private-sector data into official datasets.
- Improve documentation and quality of prominent AI datasets to reduce the number of situations with biased results.
- Provide funding from core federal agencies to close both the digital divide and the data divide.

- Direct federal agencies to update or establish data strategies to ensure data collection is integrated into diverse communities.
- Amend the Federal Data Strategy (FDS) to identify data divides and direct agency action.
- Establish a bipartisan federal commission to study the data divide.

UNDERSTANDING THE DATA DIVIDE

The data divide refers to the gap between individuals and communities that have adequate data collected and used about them and those who do not. As the data-driven economy and society continues to develop, those without sufficient data will find that many services work less effectively for them. Consider health care: Patients without detailed electronic health records (EHRs) will not benefit from health analytics and thus may receive suboptimal care; patients without wearable medical devices will not receive alerts about health abnormalities and thus may not receive life-saving treatments; and patients that are part of groups underrepresented in genetic databases will not benefit from precision medicine. In short, the data divide means that not only will some data-driven services not work for certain people and groups, but data-driven decisions may even be wrong or harmful for them. Without action, a data-driven world will leave some of these individuals and communities behind.

Advances in technology have always created this possibility. Sometimes this is because new technologies are “luxury” items that only upper-income people can afford. Private jets, luxury automobiles, high-end entertainment systems, and other similar products fit into this category. Other times, the technologies are, or at least should be, mass items everyone is able to access, especially as the technologies diffuse through society. Think cell phones, air-conditioning, electricity, and household appliances.

Addressing the data divide requires a number of considerations of the ideal versus the practical—and not all data divides require equal effort to reduce, nor should they have equal prioritization. In a data-driven world, data equity is what matters, meaning baseline data systems have representative analytics going in and accurate, actionable insights coming out.

When data divides impact public goods, closing those gaps should be a top priority. The objective of government is to serve all its constituents. Instances such as government surveys underreaching certain communities or city governments distributing smart sensors unequally should be a high priority. Similarly, the United States should strive for universality for system data, as the ramifications of data divides pose the greatest risk of harm in these areas. For example, a data-driven education system should exist in all communities, regardless of income level. In other areas, such as with wearable smart devices, getting total equality may be costly and ultimately detract from the wider goal.

Electrification initially created a difference in living standards between urban and rural areas. The rise of automobiles created communities with car dependency that made it hard for those without a vehicle to live and work. And the development of the Internet created the digital divide wherein differences in access to information technology (IT), Internet use, and digital skills can create significant disadvantages.

The data divide is similar to these past technological divisions in that the consequence for some is they will have inequitable access to the benefits of the data economy. However, the causes of the data divide, and thus the solutions to it, are unique because data is unlike other goods. Electricity and Internet access are fungible goods. For most consumers, the electricity and Internet service they receive from one provider is identical to that of another. And while someone may greatly prefer a Rolls-Royce over a Kia, in so far as providing basic transportation, vehicles are mostly interchangeable too. But data is not fungible; one set of bits is not the same as another. Giving an individual someone else's data, such as another person's health records, does not help them address their own health-care needs.

These differences matter when it comes to providing solutions. For example, policymakers have sought to close the digital divide by working to increase access and affordability of Internet services and computers. But addressing the data divide will require new thinking about how to collect and make available for use the unique datasets necessary for different individuals and groups to thrive in the data economy.

The data divide is the result of incomplete or missing datasets, including those that are not sufficiently representative of certain populations, cannot be disaggregated to address the needs of different populations, do not address relevant issues, or are not of sufficient quality for a given purpose. Individuals produce a vast amount of data from many different sources, including Internet of Things (IoT)-connected sensors, wearable devices, and payment transactions. Therefore, data divides may have a serious impact on individuals obtaining many of the benefits of using data in sectors such as financial services, environmental monitoring, education, and health care. As services increasingly rely on data in the digital economy, disparities will continue to arise between the data haves and the data have nots.

Moreover, the data divide pertains not only to the quantitative information within datasets themselves, but also to data collection methods. For example, government surveys provide the core collection method for all federal statistics. But data divides also relate to system data collection, such as with health care and credit data. Data collection can also refer to sensor data collection through IoT-connected devices such as smart appliances or neighborhood security systems. Addressing the data divide means not only addressing representation issues within statistical surveys but in data collection in all these emerging areas.

In some cases, the social and economic inequalities that may result from this imbalance in data collection and use will be so extreme that some people and groups may experience "data poverty," wherein a dearth of information on oneself and one's community has a significantly negative impact on one's quality of life.[1] Data poverty is a somewhat of a phenomenon because what is considered sufficient data changes over time as technologies mature and evolve.

There are two important aspects to the data divide that each matter at both the individual and group levels: data representativeness and data availability. Data quality, whose many dimensions include accuracy y, timeliness, precision, and completeness, impacts both data representativeness and data availability. Poor data quality is an important contributing factor to why data divides may persist.

Figure 1: Impact of representativeness and availability of data at the individual and group levels

	Individual level	Group level
Data representativeness	Datasets are not accurate about an individual	Datasets do not accurately reflect group
Data availability	Individuals cannot use services that require data about themselves	Groups cannot use services that require data about them

DATA REPRESENTATIVENESS

Data representativeness refers to the extent to which a given dataset sufficiently describes the characteristics of a larger population. A nonrepresentative dataset either excludes a certain population outright or excludes details about that population. For example, assume figure 2 shows a representative dataset listing U.S. voters and their political parties. Within that figure, table 2 shows the scenario in which a group is completely excluded from the dataset, as voters not affiliated with the two major political parties are excluded outright. Table 3 shows how a dataset may contain incomplete information about a certain group: Voters for minority political parties are listed, but the details about their specific party affiliation are left out. Both cases can contribute to individuals or groups being invisible in the data.

Figure 2: Example of nonrepresentative datasets

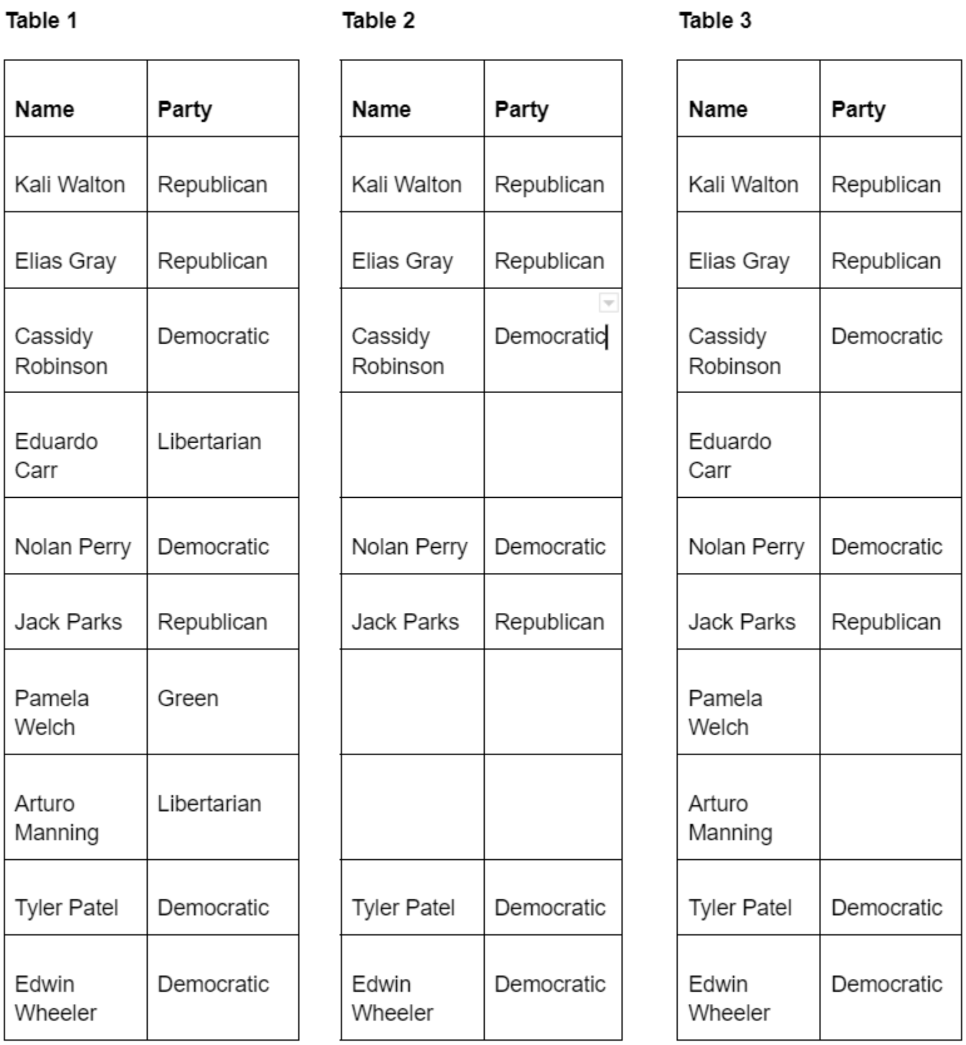


Table 1

Name	Party
Kali Walton	Republican
Elias Gray	Republican
Cassidy Robinson	Democratic
Eduardo Carr	Libertarian
Nolan Perry	Democratic
Jack Parks	Republican
Pamela Welch	Green
Arturo Manning	Libertarian
Tyler Patel	Democratic
Edwin Wheeler	Democratic

Table 2

Name	Party
Kali Walton	Republican
Elias Gray	Republican
Cassidy Robinson	Democratic
Nolan Perry	Democratic
Jack Parks	Republican
Tyler Patel	Democratic
Edwin Wheeler	Democratic

Table 3

Name	Party
Kali Walton	Republican
Elias Gray	Republican
Cassidy Robinson	Democratic
Eduardo Carr	
Nolan Perry	Democratic
Jack Parks	Republican
Pamela Welch	
Arturo Manning	
Tyler Patel	Democratic
Edwin Wheeler	Democratic

DATA AVAILABILITY

Data availability refers to the extent to which one can use sufficient data for a given purpose. Data may not be available for many reasons, including because it was never collected, it was collected but not retained, or it was collected and retained but cannot be used for some other reason, such as a technical or legal restriction. A lack of data availability means data cannot be used for a given application or service.

Data availability depends on the intended uses of data, and different individuals and groups have different data needs. For example, communities in the southeast of the United States may have more interest in detecting hurricanes and therefore require more data from storm-tide sensors whereas communities on the West Coast may be more interested in detecting earthquakes and therefore require more data from ground motion sensors.

Data availability is essential to both individual and group “data wealth.” In the digital economy, data is a factor of production—much like land, labor, and capital—enabling enterprises to make better decisions and produce goods and services more efficiently.[2] Data is also a necessary resource, much like social capital, individuals can leverage to make better decisions and utilize services to lead happier, healthier, and more-satisfying lives. Data availability gives greater agency and opportunity to individuals and communities.

THE DATA DIVIDE AFFECTS INDIVIDUALS AND GROUPS

At the individual level, a lack of representativeness means datasets do not accurately represent a person. A lack of availability means a person cannot use services that require data about themselves. Similarly, a lack of representativeness at the group level means datasets do not accurately reflect the group. A lack of availability means the group cannot use services that require data about them. In general, data divides exist on a spectrum, with some Americans falling on one end where the data collected or used about them includes them entirely or the other, where it completely excludes them. But for most individuals and groups, data divides exist on a gradient wherein the data collected or used about them includes some of their information, but not all, and the extent to which they may suffer from those inequalities varies. In the example from figure 2, insufficient representation leads to insufficient information about individual voters and minority political parties.

To see these elements of the data divide in practice, consider a city that decides to install allergen monitoring stations in four of its five neighborhoods to collect real-time data about grass and weed pollen density. The city develops an app so its citizens can predict whether they will experience allergies that day based on data from their EHRs. At the individual level, a data divide based on availability emerges if an individual does not have data about their allergies in their EHRs. Without this data, they cannot make use of the mobile app. A data divide based on a lack of representative data also emerges at the individual level for people living in the neighborhood without a monitoring station. They will likely receive inaccurate predictions for pollen levels in their area and be forced to make uninformed decisions about time spent outdoors. At the group level, there is also a data availability problem. Since the city is only collecting data about grass and weed pollen, there is no data for people who suffer from tree pollen allergies that cannot make effective use of the mobile app. Similarly, the data is not representative of all types of pollen allergies. If city planners use the data to predict the impact of landscaping decisions, they may overlook the impact on the excluded populations. In this example, closing the data divide would require several steps, including 1) ensuring everyone with pollen allergies has this information recorded in their EHRs, 2) collecting data on tree pollen density, and 3) collecting data from all the city's neighborhoods.

Table 4: An example of the data divide

	Representativeness	Availability
Individual	Inaccurate predictions for an individual living in an unmonitored neighborhood	No prediction for a person who doesn't have an electronic health record with allergy information
Group	Inaccurate prediction of the impact of city planning on tree pollen allergies	No prediction for people with tree pollen allergies

TYPES OF DATA DIVIDES

Data divides generally fall into one of three categories: system data, geographic data, or demographic data. In each of these categories, data divides may be the result of various causes, including underrepresentation in datasets, invisibility in datasets because the data cannot be disaggregated, poor quality data, and insufficient data for a given purpose.

Underrepresentation in datasets occurs when an individual or community is inadequately reflected in data, be it from counting in federal statistics or geographic placement of data collection technologies. Invisibility in datasets occurs when data is representative but cannot be disaggregated, or separated into sub-categories, thereby obscuring issues only specific groups experience. Poor data quality makes it so that available data cannot be put to practical use. Lastly, a lack of data equity means an individual's or group's unique data needs are not being met, as either may require data specific to their circumstances.

System data divides result when there is insufficient data collected in key data systems needed in areas such as education, transportation, health care, financial services, and the environment(e.g., credit reporting agencies may not collect data about certain financial activities, making it harder for certain individuals to obtain credit). Geographic data divides result when there is insufficient data about particular places (e.g., rural areas may lack the advanced sensor networks available in urban areas). Demographic data divides result when there is insufficient data about certain populations (e.g., a survey may not be representative of a particular race or age group, or obtaining wearable device data may only be affordable for high-income users).

Table 5: Types of data divides, with types, examples, and impact

Category	Type	Example	Impact
Demographic	Gender	Government vehicle crash tests only require collecting data about “male” dummies result in safety datasets that underrepresent women.	Vehicles designed for safety of average-sized male put women at a greater risk of serious injury or death.
Demographic	Race	Census undercounts American Indians and Native Alaskans.	Insufficient and ineffective resource allocation exists for broadband on tribal reservations.
Demographic	Disability	Many communities do not have data on the accessibility of sidewalks.	Navigation apps do not work as effectively for people with mobility-related disabilities.
Demographic	Income	Lower-income Americans have less personal health data because they have lower rates of ownership of smart watches and fitness trackers.	Lower-income Americans do not benefit from potentially life-saving real-time health monitoring.
Geographic	Urban/rural	People living in rural areas produce less crowdsourced data due to lower device usage.	Apps dependent on crowdsourced data may be less accurate.
Geographic	Income	Cities deploy fewer sensors to lower-income areas.	Residents in low-income areas do not have updated public health information.
System	Education	Many states do not have systems to collect and store robust, longitudinal data about students.	Residents and leaders cannot measure the link between educational programs and outcomes.
System	Health care	Some patients have either no electronic health records or incomplete ones.	Patients with no or low-quality electronic health records receive less-accurate diagnoses and treatments.
System	Financial services	Credit agencies do not collect and use alternative credit data, such as on-time rental payments.	Consumers without a traditional credit history have unnecessarily low and misleading credit scores.

Category	Type	Example	Impact
System	Environment	Federal legislation sets minimum environmental monitoring standards based on population size, resulting in no or low data collection in certain high-risk communities with smaller populations.	Certain high-risk communities lack important environmental information about their surroundings.

SYSTEM DATA

Many systems collect and store data needed for key services, including education, health care, public health, local government services, transportation, financial services, and others. Data divides can occur when there is insufficient data in these systems for certain individuals or places. There are a few ways this may occur. First, some system data may not exist or may be poor quality because of weak or nonexistent data infrastructure. Second, some system data may exist but individuals and communities face obstacles to accessing it. Lastly, some system data may be high quality but insufficient deployment of data-driven technologies hinders its use.

Education

Schools cannot take full advantage of data-driven technologies when they lack the systems necessary to collect and utilize high-quality data. Moreover, when teachers, students, and families lack robust educational data, they are forced to make decisions about enrollment or interventions based on incomplete or inaccurate information.

A strong educational data system links student data from early childhood through post-secondary education and beyond, and relies on substantial, responsive data warehouses.[3] Educational data warehouses store data from schools' learning management systems and student information systems. This data includes classroom performance, disciplinary actions, attendance, academic history, health records, and even social services casework.[4] Schools often lack the specialized systems needed to analyze these data sources, leaving data in different formats siloed and inaccessible to many key stakeholders. These systems, known as Statewide Longitudinal Data Systems (SLDSs), store data on education from early childhood to the workforce (P-20W). P-20W data allows for policymakers to perform high-level analysis about existing disparities and opportunity gaps. For educators, this data helps identify students who need additional support, while families can use this data to advocate for their students and make enrollment decisions at all stages of their educational journey.

While all U.S. states and territories apart from New Mexico have received federal funding to build an SLDS, only 17 states and the District of Columbia had built a fully linked P-20W system as of 2017.[5] And yet, only with robust, longitudinal data can stakeholders answer critical questions about equitable access to learning opportunities, make connections between different situations and educational performance, and identify trends with post-secondary education. For example, students and families rely on data about themselves and their communities in order to make consequential decisions about where to enroll for high school. Without access to robust educational data, they will be ill-equipped to make an informed decision about whether to go the traditional high school route or enroll in a Career Technical Education program. Similarly, parents need information about children with similar backgrounds in order to weigh early education options. Connecting health outcomes with educational data is another example of the importance of responsive data systems. School administrators can look at historical health trends for their community and how it impacts performance. Based on the results, they might order vision screening for a given school to see if intervening early can impact a student's attention and engagement in the classroom.[6]

Health Care

From pharmaceutical development to delivery of care to precision medicine and public health monitoring, health data is vitally important to both individuals and communities. Researchers use health-care data to measure the efficacy of new treatments and devices, as greater data production and collection leads to better understanding of biological, environmental, and social determinants of health.[7] Older or underfunded data infrastructure restricts patients, providers, and researchers in their understanding of individual and community health.

Health care generally lags behind other sectors in updating technologies for the digital age.[8] Robust health data infrastructure needs to be able to extract info from registries, administrative data, EHRs, and even patient-owned wearable devices. Otherwise, stakeholders may find themselves "data-rich and information-poor," unable to glean insights from an abundance of fragmented data.[9] In a data-driven world, health-care providers need to have a whole-person view. With clean, standardized data, providers can get insights about which high-risk patients really need vaccines, or who might need preventative care based on their health history.

Electronic Health Records

As of January 2022, 90 percent of nonfederal acute care hospitals nationally use certified EHR technology.[10] EHRs improve health care and lower health-care costs, yet only 55 percent of hospitals use these systems to exchange patient data, and 73 percent have challenges exchanging patient information across different EHR systems.[11] EHRs provide flexible patient access, reduce errors in medical records, reduce redundant testing and diagnostic procedures, aid better longitudinal tracking, and improve research data.[12] While some states, such as California and North Carolina, have mandates to ensure data sharing by different health systems, most states leave infrastructure and related issues up to the individual state health system.[13] Incomplete EHRs mean less-accurate diagnoses and treatment for patients and can also mean incomplete datasets for public health research.

Patient-Generated Health Data

In addition to divides between those who have updated and interoperable EHRs and those who lack such records, EHRs can create data divides when they are unable to adequately account for patient-generated health data (PGHD). PGHD, or data created by patients themselves using anything from personal medical devices to fitness trackers to handwritten lists, can supplement clinical data gaps.[14] Some providers view PGHD as the key to whole-person health perspectives—as the data is produced in real time by the individual patient—and can be important both for better diagnosing, monitoring, and individual advocacy.[15] PGHD often comes from wearable devices (which face their own use divides) but has the potential to facilitate delivery of care to patients, particularly for those with diabetes or hypertension.[16] For example, a child with type 1 diabetes needs routine monitoring of their blood sugar levels to ensure that their mood, energy, and overall well-being remain stable. They may wear a device to monitor their glucose levels and alert them when their levels leave the target range. Those children that can access and use a data-sharing wearable device may be able to participate in activities they otherwise would not be able to, such as attending sleepaway camp, because their caregivers and health-care providers can monitor their glucose levels remotely.

At present, PGHD remains challenging for providers to integrate into their clinical practices, as issues with interoperability, data security, privacy, and standardization all contribute to the underutilization of PGHD in practice. Responding to more types of data using different formats and measurement standards can create a large burden for providers. But as systems improve and become more efficient, that burden will reduce and the benefit to patients will continue to increase.

Prescription Drug Monitoring

The opioid epidemic has had far-reaching impacts, from widespread loss of life to severe social and economic costs. Although many factors contributed to the making of this epidemic, a lack of high-quality data and ineffective use of data-driven technology has let the crisis continue growing.[17] This dearth of data has cost many lives and damaged communities across the country. Without better data, authorities cannot take effective action in combatting the issue. One way states have tried to alleviate the issue is through prescription drug monitoring programs (PDMPs).

PDMPs are databases state governments run to track the prescribing and dispensing of controlled substances.[18] These programs use data to identify which doctors and pharmacists are prescribing controlled substances and which patients are receiving these medications, with the goal of limiting abuse and misuse. As of 2017, 54 percent of prescriptions in Oregon's PDMP data system were for opioids.[19]

While data divides for PDMPs can be related to income, gender, and race, they also exist in terms of the availability and efficacy of the information system itself. All 50 states and the District of Columbia have PDMPs, although specific reporting requirements vary by state.[20] These different requirements create a number of issues concerning interoperability and availability of data. Each state PDMP is highly variable. As a result, data sharing may be restricted, either due to a lack of interoperable data or legal restrictions that limit the effectiveness of the database.[21] For example, states differ in how often providers must report data.[22] Others limit interstate data sharing and how long the state government can hold the data.

Public Health

COVID-19 created a public health data crisis in the United States. Numerous divides emerged in terms of who and what information was counted by national statistical offices, and subsequently who had access to the necessary resources to combat the crisis. Public health has been underfunded in the United States for decades.[23] The communities with access to data collection and monitoring are more likely to have positive outcomes, as they are equipped with the best information needed for risk calculus when deciding about the safety of various activities and opportunities.

At a time when the need for evidence-based decision-making is greatest, public health data infrastructure has been greatly lacking. The National Center for Health Statistics (NCHS), a statistical office under the Centers for Disease Control and Prevention (CDC) responsible for data collection of vital statistics, interview surveys, examination surveys, and health-care provider surveys, has lost funding every year since 2009.[24] A lack of funding weakens the support for the infrastructure necessary for monitoring the impacts of COVID-19. NCHS has subsequently needed to reduce sample sizes for its survey programs. This leads to even less information about certain groups, such as those already small in size, (e.g., Native Americans). It also has led to a dearth of information about nursing homes and assisted living and extended care facilities. A lack of data collection on these communities has obscured the severity of staffing shortages in care facilities and the health status of these residents.

Additionally, national data on mental health indicators regarding the mental, emotional, and behavioral state of children is limited due to the technical and legal complexities related to the collection of children's data.[25] No comprehensive system for monitoring children's mental health exists in the United States, despite it being a key public health concern.[26] Data on mental health indicators can show where public health strategies are needed and give better insight into different interventions and their efficacy.

Financial Services

The financial services industry revolves around data, but credit agencies often lack the necessary data infrastructure to collect and score individuals based on novel forms of credit data. When determining whether someone qualifies for a loan or mortgage, or can even obtain certain services, agencies typically look at "traditional" credit data, or information about a consumer's financial borrowing and repayment history from traditional financial institutions, and leave out "alternative" data, or information apart from repayment history that paints a more accurate picture of their financial status.[27] Alternative credit data can include rent or utility payments, cell phone bills, or even cash-flow data.

Because credit reporting bureaus, such as Experian, Equifax, and TransUnion, do not account for alternative types of data in their systems, many people in the United States may have unnecessarily low or inaccurate credit scores. Credit bureaus face not only technical but legal restrictions on the use of alternative data. As a result, their information systems are ill equipped to incorporate data outside the traditional scope. For example, the credit system cannot presently incorporate on-time rental payment data. The Federal Privacy Act of 1974 and other privacy laws prevent the U.S. Department of Housing and Urban Development (HUD) from reporting on-time rental payments without prior consent.[28] A 2020 report from HUD finds that full-file reporting of rental history benefits public housing tenants and enhances the accuracy of their credit scores.[29]

Environmental Data

The extent of environmental monitoring varies widely in the United States depending on a number of factors, ranging from the industries involved to the geographic remoteness of a location to the relative vulnerability of a community. Data on air and water quality, soil contamination, microplastic pollution, and more plays a key role in determining how the environment impacts human health and risk. While national environmental monitoring technology the Environmental Protection Agency (EPA) uses is accurate and high quality, insufficient deployment leaves certain areas without information about their surroundings. Without widespread environmental monitoring, some communities must rely on anecdotal evidence to articulate or provide the extent of these risks for regulatory purposes.

Environmental monitoring systems sometimes leave out high-risk areas, such as those with heavy industry or high levels of traffic congestion and pollution, due to the high costs associated with monitoring stations.[30] For example, a 2019 refinery explosion in south Philadelphia did not register as consequential on the EPA's federal air quality index even as the explosion released 5,000 lbs. of toxic hydrofluoric acid into the air.[31] Although EPA governs 3,900 air quality monitoring devices nationwide, the network's geographic footprint remains limited. More than 120 million Americans live in counties where EPA does not monitor small-particle pollution (measuring less than 2.5 microns). This under-monitoring often occurs due to restrictions within environmental legislation that set minimum monitoring standards based on population size.[32] These standards restrict environmental agencies in their ability to monitor communities with smaller populations, and subsequently result in monitoring stations being distributed unevenly.

GEOGRAPHIC DATA

Data divides can also emerge based on geographic location, such as between urban and rural areas and high-income and low-income communities. Different geographic areas receive different investment in data-driven technologies, creating divides on numerous scales: neighborhood, city, state, and region. Depending on where they reside, a person may or may not have access to information about climate, pollution, public safety, transportation, agriculture, crime, and more.

Urban vs. Rural Areas

Rural areas have different resources and economic potential than do urban areas.[33] As such, treating rural and urban areas the same when it comes to data-driven technology is ineffective. A city of 3,000 might differ greatly from a suburban town of 50,000. Moreover, communities increasingly have opportunities to leverage "smart city" technology—low-cost sensors, wireless communication systems, automated devices, and advanced data analytics—to address key challenges such as traffic congestion, crime, and pollution, as well as improve the quality of government services and decision-making.[34] But smart city investments often make more sense in urban areas with high population density and more resources than in rural areas.[35] While urban area housing might be primarily in apartment blocks, in rural areas, predominantly single-family housing means urban solutions won't necessary apply. Similarly, rural residents face different challenges with labor, crime, education, health care, and the environment. Moreover, the digital capacity of these areas greatly differs from that of urban centers.[36]

People living in rural areas have lower percentages of device usage and participation in online activities.[37] As a result, there is less data generated about rural areas from these devices and activities. Limited data may mean that some services that depend on crowdsourced data, such as real-time road conditions, are less accurate in rural areas. Similarly, if people in rural areas generate less data, they will be underrepresented in aggregated datasets.

High-Income vs. Low-Income Communities

Geographic data divides also emerge based on an area's median income. For example, higher-income areas may have more resources to invest in data collection and sensor technology compared with lower-income areas.

City governments increasingly rely on smart technologies to manage and improve urban infrastructure.[38] The combination of physical and digital infrastructure in the form of sensors and other IoT-connected devices can help cities collect data about historically impoverished neighborhoods, which are often the hardest to reach. Smart infrastructure can be anything from sensors alerting waste management workers when bins are full and need collection to traffic control systems that can self-adjust based on the number of vehicles on the road.[39] These smart technologies enable data-driven decision-making by city officials. Moreover, when placed in neighborhoods historically neglected by city government, they can give residents better information about their communities and connect them with city services.[40]

These smart-city tools are often distributed unequally, which can create a nonrepresentative sample. For example, smart technology embedded into public transportation systems collect information about users but leave out information about those who live in transit deserts or cannot afford to use it. Without targeted investment, low-income neighborhoods typically lack the high-tech sensors seen in more-affluent areas and may exist as areas with a scarcity of data collection known as "sensor deserts."[41]

Sensor deserts, particularly in lower-income areas, can mask wider public or environmental health concerns.[42] In Chicago, the city's Array of Things (AoT) project exemplifies distributional divides in sensor technologies between high- and low-income neighborhoods. Array of Things is built on a network of various nodes, aiming to reach 500 in total across the city. As of 2020, the city had 126 nodes containing different sensor types for recording temperature, light, pollutants, and more.[43] Still, the sensor placement is highly correlated with income. Census tracts aligning with the highest-median-income decile and the lowest decile likely have sensor placement, and tracts outside these extremes likely lack sensors altogether.

Some data collection may also be focused on lower-income neighborhoods. For example, cities have begun to use gunshot-detection systems—a network of acoustic sensors throughout a city to alert police whenever it detects gun shots. However, cities may only place these systems in certain neighborhoods. For example, the technology covers approximately 25 percent of the total area of in Washington, D.C., with most of this in lower-income neighborhoods.[44]

Geospatial Data

3D elevation data in the United States remains incomplete. As of 2021, 84 percent of the continental United States had been mapped as part of the 3D Elevation Program managed by the U.S. Geological Survey's National Geospatial Program.[45] This program aims to collect high-quality topographical data of critical importance to safety, environmental health, and infrastructure projects. Those areas without this type of mapping coverage are less likely to be able to leverage analytics for geological resource assessment, precision agriculture, infrastructure management, or timely wildfire response.[46]

DEMOGRAPHIC DATA

One type of data divide is based on such factors as gender, race, age, disability, and income: demographic data. Data may not accurately represent certain demographics, either because it does not sufficiently sample a particular demographic (e.g., a survey dataset excluding elderly Americans) or because certain demographic details are excluded from the dataset (e.g., a survey dataset not including age). Data may also not exist to address the specific needs of certain demographics.

Racial and Ethnic Data

Gaps in data exist on multiple dimensions for different racial and ethnic groups, as data divides can occur when datasets do not include information on race and ethnicity, certain racial or ethnic groups are underrepresented in datasets, or there is insufficient data available to understand key race-related issues. This demographic data on its own illustrates instances when common correlations, such as that of race and income, diverge.

The Office of Management and Budget (OMB) sets standards for race and ethnicity data collected by federal agencies.[47] It uses five categories for race and one category for ethnicity as the minimum requirements for collecting. For race, federal agencies must collect data for the categories of American Indian/Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and white. For ethnicity, agencies must collect at minimum whether a person identifies as either Hispanic/Latino or not Hispanic/Latino.

These categories are broad and may miss important nuances. For example, combining Asians into one aggregate group might show that, as a group, Asians are affluent and educated. But disaggregating Asians by nationality reveals major economic diversity.[48] Similarly, the minimum OMB standards for race do not account for the experiences of those with Middle Eastern or North African origin—and people with these backgrounds often question how to categorize themselves.[49]

American Indian and Alaska Native (AI/AN) people frequently face data collection and reporting issues in the United States. Making up just 1.7 percent of the U.S. population, AI/AN communities lack a great deal of information about their own ethnic group.[50] With such a small sample size and risk of large margin of error, this group is often reduced to an asterisk in datasets about schools, health-care facilities, various professions, and even the military.[51] This reduction may be due to the costs associated with full population surveys, or necessary in order to protect the confidentiality of small communities. Moreover, such omission disproportionately affects those AI/AN people living on or near reservations.[52]

Widespread undercounting of these communities is a phenomenon with historical roots continuing to the present. When the decennial census was done by direct enumerators, AI/AN people were often misclassified as white due to strict requirements that enumerators only record a person as AI/AN if they were formally enrolled in a tribe. Because federal policy relocated many AI/AN people to different urban areas around the country, a number of them lost their connection to a specific tribe.[53] Although the census now allows for self-identification, data quality remains an issue for these groups wherein many members identify as multiracial.

The ramifications of this gap in federal statistics can impact federal funding, with less resources allocated than are needed. For example, the Federal Communications Commission (FCC) and National Telecommunications and Information Administration (NTIA) support a number of programs to bring broadband access to native lands. NTIA's Tribal Broadband Connectivity program allocates billions of dollars to increase digital access in tribal communities, yet poor data quality about the native population means government officials cannot effectively grasp the scope of the problem.[54] In addition, tribal areas are often left out of geospatial mapping performed by broadband providers, which means getting an accurate picture of connectivity is nearly impossible.

Racial and ethnic data divides are also consequential to public health outcomes in which misclassification by the government on surveys and individual physicians on health records alike can create hidden disparities and lead to inaccurate explanations. Yet, public health data consistently omits or misrepresents racial and ethnic identities.[55] Throughout the COVID-19 pandemic, states continually failed to meet OMB race and ethnicity reporting guidelines. For example, in the early months of the pandemic, only 20 states reported disaggregated data for Native Hawaiian and Pacific Islanders.[56] Some states combined the category with Asian, a combination that has not been used by the federal government in over two decades. Others did not collect this data at all.[57] Communities predominantly made up of this ethnic group could not gauge their relative risk level at a time when disaggregated health information was of the utmost importance. Better data on specific racial and ethnic groups can also highlight key disparities, such as the fact that non-Hispanic American Indian or Alaska Native people are 3.1 times more likely to be hospitalized due to COVID-19 than are non-Hispanic white people.[58]

The majority of participants in genomics are of European descent, limiting the utility of data-driven precision-medicine treatments for Americans outside the group.[59] Although data-driven drug development promises to meet the unique needs of individuals and their backgrounds, treatments for certain diseases will be rendered less effective if there is not enough data collected about people with those conditions.[60]

Insufficient data for certain racial and ethnic identities also exists in language translation. Low-resource languages, or those with low levels of data availability, create increased difficulty for natural language processing.[61] Speakers of these languages have fewer digital tools available to them. For example, in several neighborhoods in New York City, significant portions of the population are limited in their proficiency in English, such as Sunset Park in Brooklyn, where according to 2019 data from the NYC Civic Engagement Commission, 50 percent of the population has limited English proficiency.[62] Hundreds of Sunset Park's residents speak low-resource languages, including Yiddish, Hindi, Bengali, Vietnamese, and Urdu (which is considered a low-resource language despite being one of the most widely spoken languages in the world).[63] In fact, New York City has tens of thousands of citizens of voting age with limited English proficiency, many of whom speak low-resource languages. The fewer digital tools these citizens can access, the more barriers they face to casting their ballot and participating in a democracy. In addition, some languages spoken by AI/AN groups, such as Yup'ik Eskimos, lack datasets altogether.[64] As a historically oral language, no machine translation tools exist for Yup'ik and many other indigenous languages. While a world in which all uncommon languages are machine translatable remains far off, language preservation is an area where data divides persist.

Income Data

Income data acts as a measure of economic well-being and provides insights into the lives of people with different incomes and their unique needs.[65] For example, low-income Americans face different barriers to housing, health care, and financial services, among other things, than do those with higher incomes. Similarly, they have more adverse health outcomes and a higher risk of mental illness and chronic diseases and lower life expectancy.[66] But data collection, particularly sensor data collection, may not reach people from different income levels, and datasets may not be representative as a result.

In federal statistics, the U.S. government acknowledges that low-income households are less likely to respond to population surveys and adjusts for nonresponse bias.[67] As a result, federal income data is largely robust and adequately captures the population by various income levels.

But sensor, IoT, and computer data collection may not be representative of different income levels. One of the main reasons is ownership of an Internet-connected device varies drastically depending on household earnings, as devices such as personal computers and smartphones provide users with instantaneous communication options and access to online information and services. Likewise, IoT technologies collect and share data about their environment and usage patterns. But ownership of certain devices such as personal computers and wearable health trackers greatly differs for higher- and lower-income users and households.

In general, higher-income users have higher adoption rates of new technologies, creating a digital use divide.[68] According to Pew Research Center, digital use among different income brackets remains markedly different.[69] Of adults with annual incomes of less than $30,000, 24 percent say they do not own a smartphone and 41 percent lack a desktop or laptop computer. When considering cell phone versus smartphone ownership, 76 percent of those earning less than $30,000 own the former, compared with 96 percent of those making more than $75,000.[70] Beyond access to standard Internet-connected devices, ownership of wearable data-driven technology varies greatly by income. Wearable IoT devices such as smart watches and other fitness trackers can collect important data about a person’s health and daily activities. Yet, Americans with higher incomes are far more likely to own this type of device.[71] Of households earning less than $75,000 per year, 31 percent report owning a smartwatch or fitness tracker, compared with just 12 percent of those making less than $30,000 per year.[72] These trackers allow users to monitor their own health data and even input additional information about themselves and their lifestyle into an app in which the data can then be shared with private sector health companies and medical professionals. Those with lower incomes are less likely to have access to this information about themselves and subsequently are unable to use such data for medical advocacy or self-improvement.

While ownership of multiple digital devices is not necessary in order to thrive in a data-driven world, increasing the number of low-income smartphone and broadband users matters much more than ownership of a secondary device or a smartwatch.

Beyond wearable devices, data divides exist for energy data and smart meter usage. Financial barriers can restrict access to this type of technology and its benefits.[73] Smart meters provide information to consumers about prices, usage patterns, and inefficiencies in energy consumption. But lower-income Americans don't always own such appliances and therefore cannot access the same energy and cost-saving benefits as can their more affluent counterparts.

Gender Data

A data divide based on gender can emerge when datasets underrepresent a certain gender or cannot be disaggregated by gender. A gender-based data divide can also occur when there is a lack of data unique to a particular gender issue, such as about a gender-related health condition or gender-based violence. The ramifications of a gender-based data divide are numerous. For example, underrepresentation of women in drug trial data can lead to regulatory approval of drugs that negatively impact women, while underrepresentation of women's bodies in crashes can result in unsafe vehicle design.[74]

In some instances, the government does not collect certain gender data.[75] For example, the Census Bureau and Bureau of Economic Analysis (BEA) do not collect gendered information about remittances or the transfer of money to a home country by a member of a diaspora.[76] These institutions largely rely on international organizations such as the World Bank and International Monetary Fund to track these transactions. BEA's own analysis came under scrutiny in 2016 for its unreliable remittance estimates.[77] If these organizations were to collect gender information about remittances, researchers could disaggregate this data by gender to give important insights about diaspora communities in the United States and their financial status. In addition, many government surveys miss entire demographic groups, such as Americans whose gender identity may be outside the traditional binary.[78] In addition, the American Community Survey (ACS) and Community Population Survey (CPS) do not explicitly collect information about single LGBTQ individuals.[79] Without this information, certain demographics are rendered invisible in datasets.

Other times, government agencies collect gender data, but in ways that reduce its utility, such as in the case of sexual violence. Unnecessary variations in methodologies and terminology for sexual violence make the data difficult to interpret and utilize.[80] For example, federal agencies use 23 different terms to describe instances of sexual violence, meaning the same type of violence could be classified in multiple ways in data.[81] In some surveys, the same act of violence may be called “rape,” “sexual assault,” “nonconsensual sexual act,” or “unwanted sexual act.” Sometimes, these variations exist to track specific populations, but other times the Government Accountability Office finds that these inconsistencies serve no substantial purpose.[82]

Sometimes, the U.S. government collects data about gender but does not make it available publicly. For example, the Census Bureau’s Survey of Income and Program Participation (SIPP) data on state-level wealth and asset ownership does not have most categories disaggregated by gender, even though the survey collects this information.

There are also instances when datasets include data disaggregated by gender but still have a relative imbalance between genders. Data-driven drug development and personalized medicine mean new treatments, improved outcomes, and lower costs for patients. But those benefits mean less to those who are not represented in the data used to drive and evaluate these new treatments. Women have been continually underrepresented in clinical trials, despite legal requirements.[83] Researchers examined over 1,400 trials and more than 300,000 participants and found major gaps related to participation in trials relating to cardiovascular disease, psychiatric disorders, and cancer—three areas of critical importance to women.[84] [85] When these groups are left out of the data, the decisions made about the safety and efficacy of treatments for patients may be biased. In some cases, women could be more likely to have an adverse reaction to a drug and may respond differently to medical devices. Some drugs have even been taken off the market due to their effects on women, which were missed in clinical trials.[86]

Having robust representation is also critical in areas such as safety testing. Women are three times more likely than men to get whiplash in a car crash because government regulators such as the National Highway Traffic Safety Commission (NHTSC) allow car manufacturers to use the average male body to test their safety features.[87] NHTSC only requires that male dummies be tested in the driver’s seat for its key crash tests.[88] Even then, the dummies are based on a standard male of the mid-1970s. A 2019 study from the University of Virginia finds that women are 73 percent more likely to be seriously injured in a car accident as a result of this underrepresentation in data.[89]

Lastly, there are certain issues for which data is lacking for a particular gender. For example, the United States ranks 46th in terms of maternal mortality rates.[90] More data is needed to understand the causes and factors that contribute to maternal mortality during and after pregnancy, such as the impact of certain lifestyle choices, preexisting health conditions, and socioeconomic background.[91] For federal statistics, the

CDC's Pregnancy Mortality Surveillance System (PMSS) collects data regarding risk factors and pregnancy-related deaths, but has largely incomplete information.[92] Despite such a centralized system, reporting is not mandatory and use of PMSS data remains limited due to CDC interpretation of the Public Health Service Act, which governs confidentiality and dictates that no data that contains identifiable information may be used without consent.[93]

Disability Data

Disability data divides occur when datasets underrepresent people with disabilities, cannot be disaggregated to understand insights about people with disabilities, or do not address the unique needs of people with different disabilities. According to the American with Disabilities Act (ADA), a disability is any physical, sensory, or cognitive impairment that makes daily activities substantially more difficult.[94] Different federal agencies and private employers may classify individuals with disabilities in different ways, because laws such as the ADA do not list the specific impairments they each cover.[95] Moreover, various social and cultural factors may impact whether a person self-identifies as having a disability, such as stigma surrounding mental illness.[96] This complexity can make data collection about people with disabilities difficult. As a result, datasets often underrepresent people with disabilities or fail to disaggregate between the specific categories of disability. Some datasets are of particular importance to the disabled population and can contribute to a data divide if left incomplete or contain low-quality data.

A number of government surveys do not collect sufficient data about disability, despite multiple federal agencies in the United States being responsible for collecting and monitoring national and sector-specific disability data. These groups include the Census Bureau, the Bureau of Labor Statistics (BLS), the National Center for Education Statistics (NCES), the National Center for Engineering Statistics (NCSES) and NCHS.[97] Unfortunately, the specific measures used by the surveys can vary and thereby make comparisons difficult, although there are efforts to standardize.[98] In addition, many other agencies collect data about certain groups (e.g., nursing home residents or inmates) where data on disability is also lacking.[99] On the international front, similar problems regarding differences in definitions and approach used to measure disability makes it difficult to compare U.S. data on disability with that from other countries.[100] The Washington Group on Disability Statistics (WG) promotes an international standard for disability measurement and encourages countries to use comparable disability data-generation and monitoring methodologies.[101] When statistical datasets do not contain information about disability, policymakers cannot understand or meet the unique needs of these individuals

There are certain datasets that may be critical for people with disabilities. Microsoft, for example, has partnered with a number of different organizations to create new datasets designed to address specific needs for people with disabilities. Microsoft and City, University of London have created a large public dataset used to train AI systems to recognize items of daily importance for people who are blind.[102] The company also partnered with Team Gleason, a nonprofit focused on improving the lives of people with ALS, to develop an open dataset of facial images of people with ALS to improve eye tracking.[103] Other datasets of specific importance to people with disabilities could be data about accessible sidewalks in a community so that people with mobility impairments are able to safely commute to school or work.[104] Data about long-term care facilities may be necessary to help emergency responders address the needs of people with disabilities during a natural disaster.

CHALLENGES TO CLOSING THE DATA DIVIDE

Data can come from a variety of sources, and includes three broad categories: sensor data, computer data, and traditional data.[105] Sensor data encompasses data from all types of data-driven devices, such as connected medical devices and IoT sensors in a community. Computer data refers to data generated by online activity, whether it be social media posts, network traffic, or electronic payments. Traditional data means data from traditional sources, such as government surveys and business records, medical records, and educational records. The data divide impacts all three categories of data.

There are many challenges that complicate closing the data divide. First, unwarranted privacy and surveillance fears and a lack of understanding about data-driven technologies fuel resistance to increased data collection. Second, many federal agencies, states, and local governments lack the financial resources needed for robust data collection and analysis. Depending on the political environment at a given moment, budgetary restraints may hinder equitable data collection and availability. Finally, the unique, decentralized nature of the U.S. statistical system creates difficulties for comparing and analyzing data.

PRIVACY

Privacy concerns hold back the development of data systems and data collection necessary for addressing the data divide. Some Americans worry about the potential misuse and abuse of their personal data and create political resistance to increased data collection. According to a Pew Research Center poll from April 2022, 81 percent of Americans think potential risks of data collection by companies outweigh the benefits, and 66 percent believe the same about government data collection.[106] Others fear the unfair data-driven decisions and object to initiatives to deploy more analytics or automated decision-making.[107]

But data holds tremendous potential to empower consumers and serve the public interest.[108] Increasing data collection and reducing the data divide will give people more benefits and greater agency. Privacy is an important facet of consumer welfare, but it is not the only one. For some groups, including some minority communities, the greatest risk is not that the government and private organizations collect too much data about them, but rather that they collect too little high-quality data.

Myths about the privacy risks of data used to train AI systems also impede data collection. Fortunately, there are various measures data holders can take to protect the privacy of personal data while also enabling its use, such as using de-identification methods and secure multi-party computation. For example, de-identification protocols enable parties to share data while reducing the risk of re-identification.[109] And federated machine learning techniques can train models across multiple distributed datasets without revealing the sensitive underlying data.[110] Increased awareness and use of these methods and techniques could help alleviate privacy concerns.

The growing patchwork of state data privacy laws also hinders data collection efforts by creating unnecessary regulatory complexity.[111] Enacting a federal data privacy law could help increase consumer control over their personal data and enhance data quality. For example, data portability allows consumers to obtain a digital copy of their personal data from an online service or app and move that information to other services and data rectification empowers consumers to correct their personal data.[112] Consumers with incomplete or nonexistent credit histories may be unable to access credit, but data portability policies can help them show, for example, their rental and utility payments, or their cash flow on prepaid card accounts to better access the financial system. Data rectification policies allow consumers to update inaccurate personal data and avoid being classified by the incorrect gender, name, or zip code.

FINANCIAL CAPACITY

Organizations often lack the financial capacity needed to implement the necessary data infrastructure and associated programs to ensure representative data collection and its subsequent accessibility. These capacity challenges take the form of required technology and labor and related start-up and sustainability investments. Although the United States has robust data collection and monitoring systems at the federal level, data systems are sometimes insufficient at lower levels of government and within smaller organizations. To be effective, these systems, whether at the local, county, state, or federal level, need to be interoperable and operate within the same lexicon and standards. Moreover, data collection from nontraditional sources comes with its own plethora of financial challenges that affect not only federal statistics but also private organizations and individual citizens and their ability to use data.

The U.S. federal government spends approximately $6 billion, or 0.06 percent of the total budget, on its primary statistical agencies. As a result, the funding needed to prevent and close data divides within federal statistical agencies alone is also high. The $6 billion figure covers 13 principal statistical agencies and 96 other statistical programs but does not cover lower levels of government, nor does it cover other types of data beyond traditional statistics, such as sensor or computer data.

At the local level, inadequate funding and resources dedicated to keeping information and systems updated can lead to gaps in data coverage and accessibility. For example, in 2021, the FBI switched to a new crime data collection system, the National Incident-Based Reporting System (NIBRS), in which statistics are gathered from local police departments.[113] While the transition has been in the works for many years, approximately 7,300 of 18,800 local law enforcement agencies in the United States did not report any crime data to the national system in 2021.[114] The two largest police departments, New York City and Los Angeles, submitted no data at all. To use the new system, nearly all staff must be retrained on data collection. Running a robust data collection system and ensuring coordination between systems requires a large amount of resources, so data strategies for entire systems, such as with crime data, can be costly to execute.

POLITICAL OBSTACLES

Data collection is a bipartisan issue, as both sides of the aisle use data as a core facet of their decision-making. Even so, specific commitments to data collection, representation, and use issues vary depending on political interests in a given period. Depending on the party in power, the committed actors, and who holds political and economic capital at a given time determines what type of data may be prioritized.

Policymakers from both parties politicize data, creating additional barriers to closing data divides. For example, conservatives often resist increased federal data collection, such as with the College Transparency Act, which would require colleges to collect and submit data about student enrollment, matriculation, and transfers disaggregated by various demographics.[115] Likewise, liberals politicize data collection relating to core partisan issues, such as immigration.[116] While the three most recent presidential administrations have shown commitments to federal data, they did so largely reflecting partisan tendencies.

Under the Obama administration (2008–2016), the U.S. government worked to ensure a greater level of government openness and civic engagement. Obama issued the "Transparency and Open Government" memorandum, declaring that open data would be a hallmark of the administration.[117] Under Obama, the United States also got its first chief information officer, chief technology officer, chief data scientist, and more. Perhaps most importantly, the Obama administration was responsible for the launch of data.gov, a government website dedicated to hosting all federal datasets in a machine-readable format.[118] Data.gov has been a critical part of a move toward data-driven decision-making in the United States and has accompanied other major policy shifts such as an updated Freedom of Information Act and the revised Digital Accountability and Transparency Act (DATA Act).[119]

The Trump administration (2016–2020) supported greater use of government data, with Trump signing into law the Foundations for Evidence-Based Policy Act of 2018, which included the OPEN Government Data Act and created new roles for chief data officers and other statistical officials.[120]. The administration also amended certain monitoring methodologies used by the federal statistical agencies. For example, it changed the EPA's methodology for determining deaths related to environmental hazards.[121] However, a widespread rejection of data did not materialize to the extent some experts originally feared.[122]

With the Biden administration, the White House has shown a commitment to generating more data in a variety of areas. The administration specifically has promoted the need for equitable data, or statistical estimates to represent the experience of underserved populations, through executive orders and the creation of the Equitable Data Working Group.[123] Unlike its predecessors, this administration has put a greater emphasis on data sharing and increasing community access to data.

PARTICIPATION

Not all individuals and groups participate similarly in data collection, with data divides often coming from a skewed representation in data. Response bias can emerge when participation in data collection does not reflect the real-world distribution of an issue.[124] Or it can arise from improper randomization of participants, gaps in participation, participant stigma, the nature of the experiment, etc.—or simply be unintentional (e.g., a survey is sent to 500 men and 500 women, but only 150 women reply compared with 495 men).

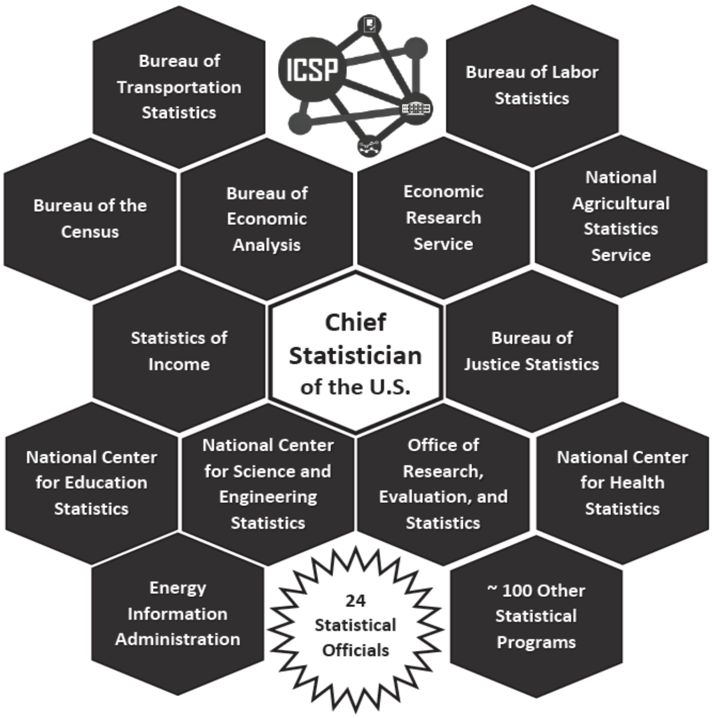
Participation gaps and biases impact all types of research, and divides can emerge when results from biased trials are incorporated into official statistics. For example, in HIV research, most participants are male and already diagnosed as HIV-positive.[125] Barriers to female participation include relative isolation, transportation, and stigma. Apart from health trials, researchers looking to gather mobility and health data from smartphone users frequently face participation bias. A study examining smartphone sensor usage finds that one's willingness to share sensor data depends on how they weigh the relative benefit and their autonomy over their data.[126]

Individuals may choose to decline participation in data collection for a variety of reasons. Sometimes, response bias and representation issues arise due to what scholars Catherine D'Ignazio and Lauren Klein describe as the "paradox of exposure," in which the same groups that stand to significantly benefit from representation in data also face the greatest risk in inclusion.[127] As a result, certain specific groups, ranging from undocumented immigrants to the ultra-rich, may be less inclined to participate in data collection. Other times, participation resistance stems from historical harms. For example, a qualitative study exploring barriers to participation in Black adults performed by a number of public health researchers finds that lingering distrust of the U.S. health-care system was a primary barrier to participation in research trials.[128] In addition, some individuals may face no risk to their livelihoods but still choose to decline to participate for other reasons.

DECENTRALIZED NATURE OF DATA COLLECTION, AGGREGATION, AND UTILIZATION

The decentralized nature of statistical collection, analysis, and dissemination in the United States makes certain groups and topics susceptible to data gaps, leading some groups to have an abundance of data about that is accessible to them and others with a dearth of information. The U.S. federal statistical system being made up of more than 125 bodies creates a number of difficulties and inefficiencies. The Office of the Chief Statistician, residing under OMB, leads the system and coordinates among all agencies, units, and programs. The chief statistician and OMB issue standards, guidelines, directives, common definitions, and monitoring methodologies. Beyond the chief statistician, there are 13 principal statistical agencies and more than 100 other statistical programs.[129] The chief statistician also chairs the Interagency Council on Statistical Policy (ICSP), which is composed of the heads of the principal agencies.[130]

Figure 3: The decentralized federal statistical system


Bureau of Transportation Statistics
ICSP
Bureau of Labor Statistics
Bureau of the Census
Bureau of Economic Analysis
Economic Research Service
National Agricultural Statistics Service
Statistics of Income
Chief Statistician of the U.S.
Bureau of Justice Statistics
National Center for Education Statistics
National Center for Science and Engineering Statistics
Office of Research, Evaluation, and Statistics
National Center for Health Statistics
Energy Information Administration
24 Statistical Officials
~ 100 Other Statistical Programs

Such a massive statistical system can impede collaboration among agencies and at times create redundancy. While sometimes redundancy can add to the reliability of collected information, other times, scholars have found that this redundancy can result in competing national estimates and create confusion and difficulty in program evaluation.[131] For example, national estimates of health insurance coverage come from both NCHS and the Census Bureau. NCHS uses the National Health Interview Survey to get insurance coverage data, while the Census Bureau uses both the Annual Survey of Social and Economic Conditions, a supplement to the CPS and ACS.

In addition to creating redundant or competing national estimates, the decentralized statistical system can hinder interagency collaboration and even coordination with state and local governments.[132] The chief statistician is responsible for collaboration and coordination oversight, but this can be difficult when statistical systems at all levels are siloed. Some agencies contract with other agencies, such as the Census Bureau undertaking surveys for other organizations.[133] But interagency collaboration can be difficult when competing organizations have differing missions and operations. Moreover, legislative restrictions can hinder data sharing even among federal and state agencies.

RECOMMENDATIONS

U.S. policymakers should address the data divide to ensure more equal opportunity in the data economy. The goal should be to speed the transition to a data-driven society while helping everyone enjoy its benefits. Policymakers should prioritize making concrete a vision of a smart, data-driven society while at the same time reducing data divides to ensure more equitable access to the new wave of data-driven innovation. Enhancing data quality, data collection methods, and dataset representation can fill critical gaps and build greater trust. Expanding funding for programs concerned with digital access to data access would help to further connect communities with data-driven technologies. Finally, policymakers should require data equity strategies to ensure the increased use of data has a balanced and fair effect.

IMPROVE FEDERAL DATA QUALITY BY DEVELOPING TARGETED OUTREACH PROGRAMS FOR UNDERREPRESENTED COMMUNITIES

To improve federal data quality, policymakers should ensure that government data collection programs develop targeted outreach programs and partner with grassroots organizations to improve response rates from underrepresented groups. The Census Bureau has historically encountered a number of difficulties in reaching certain groups, whether due to community size, language barriers, or digital access, all of which affect response rates and participation.[134] By bolstering efforts to reach these groups through intensive outreach, national statistical collection will include more communities.[135]

IMPROVE DATA QUALITY FOR NON-GOVERNMENT DATA

There are a number of ways that government can improve data quality for non-government data. Policymakers can encourage the private sector to adhere to federal standards when collecting demographic data. Government agencies should also convene stakeholders to develop best practices for collecting and sharing industry data in their respective sectors. Here, the United States could look to the United Kingdom, whose data strategy strengthens support for improved data sharing from the private sector by accelerating existing mechanisms such as “data trusts.”[136] Data trusts are legal structures that act as independent intermediaries to encourage businesses to collect and share data responsibly and can help existing and new high-quality datasets become a more widely available resource.[137]

The U.S. government can also encourage firms collecting and selling data to improve their practices by setting data quality standards for data they are willing purchase. Since many government agencies purchase data from the private sector, setting data quality standards would help incentivize compliance.

ENSURE COMPARABLE DATA COLLECTION AND MONITORING METHODOLOGIES AMONG GOVERNMENT AND CIVIL SOCIETY

While many groups collect information on data such as gender and poverty, physical and sexual harassment, and access to various financial services, they often use differing methodologies, which makes comparison and getting consistent information difficult. This dissonance also occurs internally within the federal government (e.g., federal agencies use different definitions for issues such as disability or sexual assault).[138] These differences hinder data interpretation and can create conflicting narratives about the extent of an issue. Policymakers should therefore ensure comparable collection and monitoring methodologies among government agencies and civil society organizations that collect the same or similar information. For instance, federal data should be purely digitized, with interoperable systems using the same lexicon and standards with strong enforcement. Interoperability will facilitate increased data sharing not only among federal agencies, but also with civil society organizations.

SUPPORT INCREASED UTILIZATION AND INCORPORATION OF CROWDSOURCED AND PRIVATE SECTOR DATA INTO OFFICIAL DATASETS

Even with a strong commitment to statistical accuracy and inclusion, government data can underrepresent or obscure various individuals and communities. Policymakers should continue their efforts to close data gaps and explore alternative approaches to data collection. Incorporating crowdsourced data and private sector data into official datasets and supporting its increased utilization helps give a real-time, specific snapshot of unique problems and fills knowledge gaps. Building this novel relationship between communities and the public sector can also increase trust in government.

Crowdsourced data adds a timely element to data collection, as participants answer surveys and questionnaires with their responses recorded in real time.[139] This type of data collection allows direct monitoring. Data is generated via surveys, SMS texts, phone calls, emails, and even social media, typically about a specific event or initiative. For example, crowdsourced data has been proven effective in natural disaster and humanitarian response.[140] These situations in which official capacity may be limited often rely on on-the-ground information or intelligence to understand the severity of a crisis. Crowdsourced data allows for faster, lower-cost collection of information so policymakers can act more efficiently and effectively.[141] In the 2015 Nepal earthquake, for example, the Humanitarian OpenStreetMap proved the best way to understand the extent of the disaster's damage.[142] Maps of infrastructure data such as roads, building footprints, amenities, land uses, and more were used to expedite aid delivery and reconstruction projects. The U.S. Department of State already supports OpenStreetMap through a citizen science grant, but more agencies should consider this type of data collection as one element of disaster response.[143]

Alternative data collection approaches also build a more participatory relationship between private citizens and the public sector. Individual citizens have greater agency in solving urban and environmental problems through efforts such as citizen science that document their lived experience in a structured manner and empowers them to advocate for themselves whenever harms arise.[144] Crowdsourced data incorporates marginalized groups that might be more reluctant to participate in official government data collection.[145] It can also highlight or refute misconceptions and act as a check to official statistics or understanding.[146]

Private sector data can also provide important insights that can supplement government data. For example, payroll services such as ADP have immediate insight into the labor market that traditional government surveys may take weeks or longer to gather. Similarly, private sector firms such as Datasembly provide real-time, hyper-local retail pricing that give more detail than the Bureau of Labor Statistics Consumer Price Index. Government agencies should consider where increased use of private sector data would add value and develop partnerships with private sector stakeholders to improve government data. As government agencies seek to close data gaps and address data-related inequalities, policymakers should consider increased support and use of alternative research and data collection methods.

IMPROVE DOCUMENTATION AND QUALITY OF PROMINENT AI DATASETS

Certain datasets are widely used for many applications, especially for training AI systems. The U.S. government has already rightly recognized the need to better document known issues with these critical data assets as well as improve them over time. and is exploring how it provides this resource. For example, the Office of Science and Technology Policy and National Science Foundation are developing a National AI Research Resource, a shared computing and data resource that aims to improve AI datasets while acknowledging the importance of equity and fairness in data.[147] In addition, OMB is exploring how to improve the quality and access of government datasets for AI. [148] But there are several areas where the federal government can maximize the effectiveness of its efforts to ensure AI researchers and developers use and annotate datasets properly and, as a result, reduce the number of situations where biased results emerge.[149]

First, OMB can create a set of best practices for dataset descriptions to ensure proper use of public datasets. Large datasets used to train machine learning algorithms often lack documentation on their known problems or limitations. As a result, researchers or developers may unknowingly use these datasets in situations where the data is ill suited for the application. Public datasets for machine learning can be found through a number of aggregator sites, such as Kaggle, specific academic projects (e.g., ImageNet), or via private organizations (e.g., Meta). Through millions of data points, researchers can train algorithms on natural language processing, computer vision, and even sentiment analysis. But datasets need comprehensive descriptions containing information about dataset representation, problems, reproducibility, and data collection efforts.

Large datasets also frequently have annotation issues. For example, ImageNet, a computer vision dataset with more than 14 million images and thousands of labels, still has a number of images that contain multiple objects wherein a single image label may not be enough.[150] This can lead to a 10 percent drop in the general accuracy of object recognition models.[151] Similarly, nearly 30 percent of Google's "GoEmotions" dataset containing over 58,000 Reddit comments have inaccurate labels.[152] Inaccurate or incomplete annotation can affect the quality of algorithmic predictions, or even lead to unethical or offensive errors, particularly when the dataset involves people.[153]

Some fields and organizations have committed to better dataset descriptions, such as computer scientists devising "bug tracking" to track and monitor errors in software testing. Tracking these defects allows researchers to manage and prioritize ameliorating these issues. The open source code repository GitHub has bug-tracking available to all users called GitHub Issues.[154] Similarly, Meta provides a great deal of information on its public datasets, with each dataset having an overview, followed by more detailed descriptions of its applications, intended use cases, data types, dataset characteristics, data collection and labeling methods, along with validation information.[155] For example, the Casual Conversations dataset provides specific information about the characteristics of subjects included in the dataset, such as the number of labels provided for gender and skin tone, and whether those labels are self-provided or human labeled.

OMB and other relevant federal agencies should support the development of best practices for dataset labeling and annotation, and aid the development of high-quality, application-specific training and validation data in sensitive and high-value contexts, such as in health care and transportation. These best practices should serve as an unbiased resource for organizations and researchers developing AI datasets.

PROVIDE FUNDING TO CLOSE BOTH THE DIGITAL DIVIDE AND DATA DIVIDE

From smart wearable devices to local environmental sensors, the use of data-driven technology has skyrocketed in recent years as policymakers recognize the value in evidence-based decision-making. But most government funding remains targeted at access to broadband and hardware. In the coming years, policymakers will need to think about how to ensure people and communities can participate in not only the Internet economy but also the data economy.[156]

A number of programs exist to address the digital divide, spearheaded by the FCC, NTIA, and other federal agencies. These programs, such as the FCC's Affordable Connectivity Fund and the Emergency Broadband Benefit aim to connect low-income households with broadband Internet access and cover some of the costs associated with Wi-Fi connection and laptop computers.[157] But these programs should not only focus on connectivity to the Internet, but also help connect people to data-driven services and technology. Inequitable deployment of critical data-driven technology such as educational data systems and environmental monitoring stations can worsen inequalities and even render the technology less effective.[158]

Policymakers should ensure that funding for data-focused initiatives prioritizes greater inclusion of underserved communities. For example, the Department of Energy should take digital inclusion into account and require citywide deployment when funding programs such as smart city energy data pools that address energy and climate goals.[159]

DIRECT FEDERAL AGENCIES TO UPDATE OR ESTABLISH DATA STRATEGIES TO ENSURE DATA COLLECTION IS INTEGRATED INTO DIVERSE COMMUNITIES

Systems that collect and store data, including education, health care, financial services, and environmental monitoring need strategies for minimizing potential data divides and maximizing the opportunity created by increased data collection. Bottom-up data collection technologies, such as smart city and IoT sensors offer many opportunities to embed intelligence into everyday items and generate data that can power insights into productivity, sustainability, and resilience.[160]

Congress should direct federal agencies such as the Department of Education, Department of Energy, CDC, Consumer Financial Protection Bureau, and EPA to update or establish data strategies to include provisions that ensure that this type of data collection is integrated into diverse communities. For example, the Department of Education's National Educational Technology Plans should include guidance on educational data collection systems to ensure that schools in lower-income and affluent areas alike have the capability to implement interoperable, longitudinal data systems.

AMEND THE FEDERAL DATA STRATEGY TO ADDRESS THE DATA DIVIDE

The Biden administration has shown a commitment to equitable data collection and access, as demonstrated through the Equitable Data Working Group.[161] These commitments support greater data disaggregation of demographic variables such as race, gender, and income. In line with the Working Group findings, OMB should amend the FDS to identify and address the data divide and, as a primary obstacle to data equity, should offer steps agencies can take to eliminate data divides related to their respective missions and constituencies. The current strategy aims to accelerate the use of data to serve the public, but largely focuses on federal statistics.[162]

Data equity strategies ensure a commitment to data collection and access that accounts for existing imbalances and gaps and sufficiently addresses the potential for bias. For example, the We All Count project for equity in data science promotes seven key steps organizations should take for data equity.[163] Steps include transparency in funding, encouraging consideration of the balance between more data and privacy, and getting better reliability with small sample sizes. Improved reliability of data and greater transparency can strengthen stakeholder trust and increase community engagement in the data collection process. Data equity strategies should include plans to close gaps, boost geographic distribution, and build more representative datasets. In line with an amended FDS, states and other federally supported organizations should create data equity strategies to ensure collection, analysis, and reporting of data has all people in mind.

ESTABLISH A BIPARTISAN FEDERAL COMMISSION TO STUDY THE DATA DIVIDE

While there are a number of steps the federal government can take in the near term to address the data divide, the issue deserves more attention. To that end, Congress should establish a bipartisan federal commission to study the data divide, including its causes and impacts, and provide recommendations on opportunities to close the data divide to promote more economic and social opportunities for all Americans.

The commission should engage a diverse group of stakeholders with the goal of improving federal statistics and data collection and supporting private sector initiatives to address critical data divides wherever a lack of data creates significant risk of exclusion in the data economy. It should also assess the causes and impact of high-priority data divides, as well as recommend tailored solutions to address these divides, including by considering how to use existing or new funding mechanisms for reducing data divides.

CONCLUSION

As this report identifies, data divides manifest in a number of ways and should be a top concern for policymakers. The data economy and data-driven innovation present a powerful opportunity to transform society for the better, but only if data collection and use are inclusive. Policymakers should work to ensure that all individuals and communities have access to high-quality data.

REFERENCES

1. Daniel Castro, "The Rise of Data Poverty in America" (Center for Data Innovation, September 2014), https://datainnovation.org/2014/09/the-rise-of-data-poverty-in-america/.
2. Daniel Castro and Travis Korte, "Data Innovation 1010" (Center for Data Innovation, November 2013), https://datainnovation.org/2013/11/data-innovation-101/.
3. Gillian Diebold and Chelsea Han, "How AI Can Improve K-12 Education in the United States" (Center for Data Innovation, April 2022), https://www2.datainnovation.org/2022-ai-education.pdf.
4. Ibid.
5. "50-State Comparison: Statewide Longitudinal Data Systems," Education Commission of the States, last modified December 14, 2021, https://www.ecs.org/state-longitudinal-data-systems/.
6. "Data Linkages Enable Individual Support and Shared Success," Data Quality Campaign, last modified June 30, 2020, https://dataqualitycampaign.org/wp-content/uploads/2020/06/DQC_Data-Linkages-Enable-Individual-Support-and-Shared-Success.pdf.
7. Daniel Castro, "The Rise of Data Poverty in America."
8. "Health Information Infrastructure," OECD, accessed June 2022, https://www.oecd.org/health/health-systems/health-data-infrastructure.htm.
9. Claudia Williams, "How to Get Health Data Infrastructure Right For This Moment of Medicaid Transformation," *Health Affairs* (January 2022), https://www.healthaffairs.org/do/10.1377/forefront.20220113.723542.
10. Ibid.
11. The MITRE Corporation, "A Robust Health Data Infrastructure" (report prepared for Agency for Healthcare Research and Quality, 2013), https://digital.ahrq.gov/ahrq-funded-projects/robust-health-data-infrastructure; Claudia Williams, "How to Get Health Data Infrastructure Right For This Moment of Medicaid Transformation."
12. The MITRE Corporation, "A Robust Health Data Infrastructure."
13. Ibid.
14. "Patient-Generated Health Data," HealthIT.gov, last modified March 21, 2018, https://www.healthit.gov/topic/scientific-initiatives/patient-generated-health-data.
15. Kea Turner et al., "Sharing patient-generated data with healthcare providers: findings from a 2019 national survey," *Journal of the American Medical Informatics Association*, 371–376. https://pubmed.ncbi.nlm.nih.gov/33180896/.
16. Ibid.
17. Joshua New, "How Data Can Help in the Fight Against the Opioid Epidemic in the United States" (Center for Data Innovation, November 2019), https://s3.amazonaws.com/www2.datainnovation.org/2019-data-opioids.pdf.

18. Daniel Castro, “Using Data to Fight the Opioid Epidemic,” *Government Technology,* June 2017, https://www.govtech.com/analysis/using-data-to-fight-the-opioid-epidemic.html.

19. Daniel Castro, “Data is Key to the Fight Against the Opioid Epidemic,” *StateTech Magazine*, November 8, 2017, https://statetechmagazine.com/article/2017/11/data-key-fight-against-opioid-epidemic-0.

20. “State PDMP Profiles and Contacts,” PDMPTTAC, accessed July 2022, https://www.pdmpassist.org/State.

21. Daniel Castro, “Using Data to Fight the Opioid Epidemic.”

22. Ibid.

23. Joseph Williams, “Pandemic Exposed a Public Health System ‘Hollowed Out’ From Lack of Funding and Neglect,” *U.S. News & World Report*, March 10, 2021, https://www.usnews.com/news/health-news/articles/2021-03-10/report-pandemic-exposed-a-public-health-system-hollowed-out-from-lack-of-funding-and-neglect.

24. “State of the Nation’s Health Data Infrastructure: Experts Weigh in Two Years into Pandemic,” *AmStatNews*, March 1, 2022, https://magazine.amstat.org/blog/2022/03/01/nchs/.

25. “Data and Statistics on Children’s Mental Health,” Centers for Disease Control and Prevention, last modified June 2022, https://www.cdc.gov/childrensmentalhealth/data.html.

26. Rebecca Bitsko et al., Mental Health Surveillance Among Children — United States, 2013–2019. MMWR Supplement 2022, https://www.cdc.gov/mmwr/volumes/71/su/su7102a1.html.

27. “What is alternative credit data?” LevelCredit, accessed July 2022, https://www.levelcredit.com/what-is-alternative-credit-data.

28. “Enterprise Income Verification (EIV) System” (Notice H 2013-06 from U.S. Department of Housing and Urban Development, March 2013), https://www.hud.gov/sites/documents/13-06HSGN.PDF.

29. Michael Turner and Patrick Walker, *Potential Impacts of Credit Reporting Public Housing Rental Payment Data* (Washington, D.C., U.S. Department of Housing and Urban Development, October 2019), https://www.huduser.gov/portal/publications/Potential-Impacts-of-Credit-Reporting.html.

30. Gillian Diebold, “Citizen Science and Crowdsourced Data Can Improve Environmental Data in the United States” (Center for Data Innovation, June 2016), https://datainnovation.org/2022/06/citizen-science-and-crowdsourced-data-can-improve-environmental-data-in-the-united-states/.

31. Tim McLaughlin, Laila Kearney, and Laura Sanicola, “Special Report: U.S. air monitors routinely miss pollution – even refinery explosions,” *Reuters*, December 1, 2020, https://www.reuters.com/article/usa-pollution-airmonitors-specialreport/special-report-u-s-air-monitors-routinely-miss-pollution-even-refinery-explosions-idUSKBN28B4RT.

32. “Why isn’t there an outdoor air monitor in my county?” Environmental Protection Agency, last modified October 4, 2021, https://www.epa.gov/outdoor-air-quality-data/why-isnt-there-outdoor-air-monitor-my-county.

33. “Reenvisioning Rural America,” Urban Institute, September 21, 2021, https://reenvisioning-rural-america.urban.org/.

34. Joshua New, Daniel Castro, and Matt Beckwith, “How National Governments Can Help Smart Cities Succeed” (Center for Data Innovation, October 2017), https://www2.datainnovation.org/2017-national-governments-smart-cities.pdf.

35. Johanna Walker et al, “Smart rural: The open data gap” (paper presented at Data for Policy 2020 conference, September 15–17, 2020), https://www.researchgate.net/publication/344281549_Smart_rural_The_open_data_gap.

36. Edward Carlson and Justin Goss, “The State of the Urban/Rural Digital Divide” (National Telecommunications and Information Administration), https://www.ntia.doc.gov/blog/2016/state-urbanrural-digital-divide.

37. Ibid.

38. Solomon Greene et al., “Technology and Equity in Cities” (Urban Institute, November 2019), https://www.urban.org/sites/default/files/publication/101360/technology_and_equity_in_cities_1.pdf.

39. “Top 5 Examples of Smart City IoT Solutions,” iotaComm, last modified November 1, 2019, https://www.iotacommunications.com/blog/smart-city-solutions-examples/.

40. Solomon Greene et al., “Technology and Equity in Cities.”

41. “Spatial inequality and the smart city,” The Alan Turing Institute, accessed June 2022, https://www.turing.ac.uk/research/research-projects/spatial-inequality-and-smart-city.

42. Homi Kharas and Jaana Remes, “Can Smart cities be equitable?” (Brookings, June 2018), https://www.brookings.edu/opinions/can-smart-cities-be-equitable/

43. Caitlin Robinson and Rachel Franklin, “The sensor desert quandary: What does it mean (not) to count in the smart city?” *Transactions (Institute of British Geographers)* (2020) 1–17, https://www.researchgate.net/publication/345311943_The_sensor_desert_quandary_What_does_it_mean_not_to_count_in_the_smart_city.

44. Christina Stacy, Yasemin Irvin-Erickson, and Emily Tiry, “The impact of gunshots on place-level business activity” (2021), https://www.researchgate.net/publication/351710596_The_impact_of_gunshots_on_place-level_business_activity.

45. “What is 3DEP?” U.S. Geological Survey, accessed July 2022, https://www.usgs.gov/3d-elevation-program/what-3dep.

46. Daniel Castro, Joshua New, and Matt Beckwith, “10 Steps Congress Can Take to Accelerate Data Innovation” (Center for Data Innovation, Mach 2017), https://www2.datainnovation.org/2017-data-innovation-agenda.pdf.

47. “Office of Management and Budget (OMB) Standards, National Institutes of Health, accessed June 2022, https://orwh.od.nih.gov/toolkit/other-relevant-federal-policies/OMB-standards.

48. Dedrick Asante-Muhammad and Sally Sim, “Racial Wealth Snapshot: Asian Americans and the Racial Wealth Divide” (National Community

Reinvestment Coalition, May 2020), https://ncrc.org/racial-wealth-snapshot-asian-americans-and-the-racial-wealth-divide/.

49. Juli Adhikari and Jocelyn Frye, "Who We Measure Matters: Connecting the Dots Among Comprehensive Data Collection, Civil Rights Enforcement, and Equality" (Center for American Progress, March 2020), https://www.americanprogress.org/article/measure-matters-connecting-dots-among-comprehensive-data-collection-civil-rights-enforcement-equality/.

50. "Profile: American Indian/Alaska Native," U.S. Department of Health and Human Services Office of Minority Health, last modified January 2022, https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=3&lvlid=62#:~:text=American%20Indians%20and%20Alaska%20Natives,of%20the%20total%20U.S.%20population.

51. "Data Disaggregation," NCAI Policy Research Center, accessed June 2022, https://www.ncai.org/policy-research-center/research-data/data.

52. "American Indians and Alaska Natives Living on Reservations Have the Highest 2020 Census Undercount," National Congress of American Indians, March 10, 2022, https://www.ncai.org/news/articles/2022/03/10/american-indians-and-alaska-natives-living-on-reservations-have-the-highest-2020-census-undercount.

53. Kimberly Huyser, "Data and Native American Identity," *Contexts* (2020), https://journals.sagepub.com/doi/full/10.1177/1536504220950395.

54. Roundtable discussion on "Closing the Digital Divide in Native Communities through Infrastructure Investment," 117th Congress (2022), https://www.indian.senate.gov/hearing/roundtable-discussion-closing-digital-divide-native-communities-through-infrastructure.

55. Geoffrey Holtzman, Neda Khoshkhoo, and Elaine Nsoesie, "The Racial Data Gap: Lack of Racial Data as a Barrier to Overcoming Structural Racism," *The American Journal of Bioethics*, no. 22 (2022), 39–42, https://www.tandfonline.com/doi/full/10.1080/15265161.2022.2027562.

56. Richard Chang, Corina Penaia, and Karla Thomas, "Count Native Hawaiian and Pacific Islanders in COVID-19 Data – It's an OMB Mandate," *Health Affairs* (2020), https://www.healthaffairs.org/do/10.1377/forefront.20200825.671245/full/.

57. Ibid.

58. "COVID-19 infections by race: What's behind the health disparities?" Mayo Clinic, last updated April 29, 2022, https://www.mayoclinic.org/diseases-conditions/coronavirus/expert-answers/coronavirus-infection-by-race/faq-20488802.

59. Joshua New, "The Promise of Data-Driven Drug Development" (Center for Data Innovation, September 2019), https://www2.datainnovation.org/2019-data-driven-drug-development.pdf.

60. Sarah Richards, "The DNA Data We Have Is Too White. Scientists Want to Fix That," *Smithsonian*, https://www.smithsonianmag.com/science-nature/gene-bank-too-white-180968884/.

61. Morris, “A Quick Guide to Low-Resource NLP,” MLOps Community website, accessed August 2022, https://mlops.community/a-quick-guide-to-low-resource-nlp/.

62. “Map of LEP and CVALEP Population by Community District,” NYC.gov, accessed August 2022, https://www1.nyc.gov/assets/civicengagement/html/CEC-Language-Profiles-Map.html.

63. Asad Khattak et al., “A Survey on sentiment analysis in Urdu: A resource-poor language,” *Egyptian Informatics Journal*, no. 22 (2021), 53–74, https://www.sciencedirect.com/science/article/pii/S1110866520301171.

64. Kevin Chavez and Christopher Liu, “Yup’ik Eskimo to English: Machine Translation Using Augmented Datasets” (Stanford University), https://cs230.stanford.edu/files_winter_2018/projects/6909884.pdf.

65. “Why We Ask Questions About…Income,” United States Census Bureau, accessed July 2022, https://www.census.gov/acs/www/about/why-we-ask-each-question/income/#:~:text=Income%20data%20measure%20the%20economic,%2C%20housing%2C%20and%20other%20assistance.

66. “Poverty,” HealthyPeople.gov, accessed July 2022, https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-health/interventions-resources/poverty.

67. Jonathan Rothbaum, “How Does the Pandemic Affect Survey Response: Using Administrative Data to Evaluate Nonresponse in the Current Population Survey Annual Social and Economic Supplement,” United States Census Bureau, https://www.census.gov/newsroom/blogs/research-matters/2020/09/pandemic-affect-survey-response.html.

68. Alexander van Deursen et al., “Digital inequalities in the Internet of Things: differences in attitudes, material access, skills, and usage” *Information, Communication, and Society*, no. 24:2 (2019), 258–276, https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1646777.

69. Emily Vogels, “Digital divide persists even as Americans with lower incomes make gains in tech adoption” (Pew Research Center, June 2021), https://www.pewresearch.org/fact-tank/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/.

70. “Mobile Fact Sheet,” Pew Research Center, April 2021, https://www.pewresearch.org/internet/fact-sheet/mobile/.

71. “The Future of Wearable Technology” (Trajectory Partnership, August 2021), https://trajectorypartnership.com/wp-content/uploads/2021/08/Wearables-August-2021-FINAL.pdf.

72. Emily Vogels, “About one-in-five Americans use a Smart watch or fitness tracker” (Pew Research Center, January 2020), https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/.

73. Kevin Stark, “Survey finds racial, income gaps in awareness about smart meters,” *Energy News Network*, May 16, 2018,

https://energynews.us/2018/05/16/survey-finds-racial-income-gaps-in-awareness-about-smart-meters/.

74. Susan Molinari and Beth Brooke, “Women are more likely to die or be injured in car crashes. There’s a simple reason why.” *Washington Post*, December 21, 2021, https://www.washingtonpost.com/opinions/2021/12/21/female-crash-test-dummies-nhtsa/.

75. Gillian Diebold, “The U.S. Government Should Better Collect Gender Data” (Center for Data Innovation, April 2022), https://datainnovation.org/2022/04/the-u-s-government-should-better-collect-gender-data/.

76. “What are remittances, how big are they, and why might BEA’s estimate differ from estimates released by other organizations?” Bureau of Economic Analysis, last modified August 9, 2022, https://www.bea.gov/taxonomy/term/591?page=1.

77. Government Accountability Office, *International Remittances: Actions Needed to Address Unreliable Official U.S. Estimate* (GAO, 2016), https://www.gao.gov/assets/gao-16-60.pdf.

78. Juli Adhikari and Jocelyn Frye, “Who We Measure Matters: Connecting the Dots Among Comprehensive Data Collection, Civil Rights Enforcement, and Equality” (Center for American Progress, March 2020), https://www.americanprogress.org/article/measure-matters-connecting-dots-among-comprehensive-data-collection-civil-rights-enforcement-equality/.

79. “American Community Survey,” LGBTData.com, accessed July 2022, http://www.lgbtdata.com/american-community-survey-acs.html.

80. Joshua New, “How Common is Sexual Assault in the United States? The Answer Depends on Who You Ask” (Center for Data Innovation, September 2016), https://datainnovation.org/2016/09/how-common-is-sexual-assault-in-the-united-states-the-answer-depends-on-who-you-ask/.

81. Government Accountability Office, *Sexual Violence Data: Actions Needed to Improve Clarity and Address Differences Across Federal Data Collection Efforts* (GAO, August 2016), https://www.gao.gov/products/gao-16-546.

82. Joshua New, “How Common is Sexual Assault in the United States? The Answer Depends on Who You Ask” (Center for Data Innovation, September 2016), https://datainnovation.org/2016/09/how-common-is-sexual-assault-in-the-united-states-the-answer-depends-on-who-you-ask/.

83. Erin Blakemore, “Women are still underrepresented in clinical trials,” *Washington Post*, June 27, 2022, https://www.washingtonpost.com/health/2022/06/27/underrepresentation-women-clinical-trials/.

84. Alexandra Sosinsky et al., “Enrollment of female participants in United States drug and device phase 1-3 clinical trials between 2016 and 2019,” ScienceDirect, https://www.sciencedirect.com/science/article/pii/S1551714422000441?via%3Dihub.

85. Joshua New, “The Promise of Data-Driven Drug Development.”

86. Anna Nowogrodzki, “Inequality in medicine,” *Nature*, October 5, 2017, https://www.nature.com/articles/550S18a.

[87] Caroline Criado Perez, *Invisible Women: Data Bias in a World Designed for Men* (Abrams, 2019).

[88] Susan Molinari and Beth Brooke, “Women are more likely to die or be injured in car crashes. There’s a simple reason why.” *Washington Post*, December 21, 2021, https://www.washingtonpost.com/opinions/2021/12/21/female-crash-test-dummies-nhtsa/.

[89] Fariss Samarrai, “Study: New Cars Are Safer, But Women Most Likely To Suffer Injury,” *UVAToday*, July 10, 2019, https://news.virginia.edu/content/study-new-cars-are-safer-women-most-likely-suffer-injury.

[90] “What Explains the United States’ Dismal Maternal Mortality Rates?” (Event at Wilson Center, November 19, 2015), https://www.wilsoncenter.org/event/what-explains-the-united-states-dismal-maternal-mortality-rates.

[91] “What factors increase the risk of maternal morbidity and mortality?” National Institute of Child Health and Human Development, last modified May 14, 2020, https://www.nichd.nih.gov/health/topics/maternal-morbidity-mortality/conditioninfo/factors.

[92] Rachel Mayer et al., “The United States Maternal Mortality Rate Will Continue to Increase Without Access to Data” *Health Affairs* (2019), https://www.healthaffairs.org/do/10.1377/forefront.20190130.92512/.

[93] “Section 308(d) of the Public Health Service Act,” CDC, accessed July 2022, https://www.cdc.gov/rdc/Data/b4/section308.pdf.

[94] “A Guide to Disability Rights Laws,” ADA.gov, last modified February 2020, https://www.ada.gov/cguide.html.

[95] Ibid.

[96] “Mental Health: Overcoming the stigma of mental illness,” Mayo Clinic, last modified May 2017, https://www.mayoclinic.org/diseases-conditions/mental-illness/in-depth/mental-health/art-20046477.

[97] Brianna Blaser and Richard Ladner, “Why is Data on Disability so Hard to Collect and Understand?” (University of Washington), https://www.washington.edu/doit/sites/default/files/atoms/files/RESPECT_2020_DisabilityData.pdf.

[98] Gina Livermore et al., *Disability Data in National Surveys* (Department of Health and Human Services & Mathematica Policy Research, August 2011), https://aspe.hhs.gov/reports/disability-data-national-surveys-0.

[99] Ibid.

[100] Mitchell Loeb, “Disability statistics: an integral but missing (and misunderstood) component of development work,” *Nordic Journal of Human Rights*, 31: 3 (2013), 306–324, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4766593/.

[101] “About the WG,” Washington Group on Disability Statistics, accessed May 2022, https://www.washingtongroup-disability.com/about/about-the-wg/.

[102] “The ORBIT (Object Recognition for Blind Image Training) Dataset,” ORBIT, accessed July 2022, https://orbit.city.ac.uk/.

[103] Jennifer Langston, “Shrinking the ‘data desert’: Inside efforts to make AI systems more inclusive of people with disabilities,” *Microsoft | The AI*

Blog, October 12, 2020, https://blogs.microsoft.com/ai/shrinking-the-data-desert/; Kimberly Adams, “Making datasets inclusive from the ground up,” *Marketplace Tech*, December 24, 2020, https://www.marketplace.org/shows/marketplace-tech/making-datasets-inclusive-from-the-ground-up/.

104. Daniel Castro, “5 Q’s With Ben Schmidt, Co-founder and CEO of Roadbotics (Center for Data Innovation, June 2022), https://datainnovation.org/2022/06/5-qs-with-ben-schmidt-co-founder-and-ceo-of-roadbotics/.

105. Daniel Castro, “The Rise of Data Poverty in America.”

106. Brooke Auxier et al., “Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information” (Pew Research Center, November 2019), https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/.

107. Aaron Smith, “Attitudes Towards Algorithmic Decision-Making” (Pew Research Center, November 2018), https://www.pewresearch.org/internet/2018/11/16/attitudes-toward-algorithmic-decision-making/.

108. Daniel Castro, “Improving Consumer Welfare with Data Portability” (Center for Data Innovation, November 2021), https://datainnovation.org/2021/11/improving-consumer-welfare-with-data-portability/.

109. Ann Cavoukian and Daniel Castro, “Setting the Record Straight: De-Identification Does Work” (ITIF, June 2014), https://itif.org/publications/2014/06/16/setting-record-straight-de-identification-does-work/.

110. Hugh Brendan McMahan et al., “Federated Learning of Deep Networks using Model Averaging,” February 2016, https://arxiv.org/abs/1602.05629v1.

111. Ashley Johnson and Daniel Castro, “Maintaining a Light-Touch Approach to Data Protection in the United States” (Information Technology and Innovation Foundation, August 2022), https://itif.org/publications/2022/08/08/maintaining-a-light-touch-approach-to-data-protection-in-the-united-states/.

112. “Article 16 GDPR: Right to Rectification,” intersoft consulting, https://gdpr-info.eu/art-16-gdpr/.

113. Weihua Li, “What Can FBI Data Say About Crime in 2021? It’s Too Unreliable to Tell” (The Marshall Project, June 2022), https://www.themarshallproject.org/2022/06/14/what-did-fbi-data-say-about-crime-in-2021-it-s-too-unreliable-to-tell.

114. Weihua Li, “FBI national crime data collection in 2021” (The Marshall Project, June 2022), https://observablehq.com/@themarshallproject/2021-fbi-national-crime-data-collection.

115. Jane Robbins, “Conservatives Should Oppose the College Transparency Act,” *Townhall*, April 21, 2021, https://townhall.com/columnists/janerobbins/2021/04/21/conservatives-should-oppose-the-college-transparency-act-n2588303.

116. Madeline Fitzgerald, “Democrats, Independents Fuel Rise in Worry about Illegal Immigration,” *U.S. News & World Report*, April 22, 2022, https://www.usnews.com/news/national-news/articles/2022-04-22/democrats-independents-fuel-rise-in-worry-about-illegal-immigration; College Transparency Act, S.839, 117th Congress (2021), https://www.congress.gov/bill/117th-congress/senate-bill/839/text.

117. “Transparency and Open Government: Memorandum for the Heads of Executive Departments and Agencies,” The White House, January 21, 2009, https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government.

118. “FACT SHEET: Data by the People, for the People — Eight Years of Progress Opening Government Data to Spur Innovation, Opportunity, & Economic Growth,” The White House, September 28, 2016, https://obamawhitehouse.archives.gov/the-press-office/2016/09/28/fact-sheet-data-people-people-eight-years-progress-opening-government.

119. Jessica Mulholland, “What Obama Did for Tech: Transparency and Open Data,” *Government Technology*, September 27. 2016, https://www.govtech.com/data/what-obama-did-for-tech-transparency-and-open-data.html.

120. Foundations for Evidence-Based Policymaking Act of 2018, H.R. 4174, 115th Congress (2017) https://www.congress.gov/bill/115th-congress/house-bill/4174.

121. Catherine Rampell, “The Trump administration’s war on statistics isn’t slowing down,” *Washington Post,* May 23, 2019, https://www.washingtonpost.com/opinions/trumps-grand-plan-to-rig-government-numbers-in-his-favor/2019/05/23/9449b740-7d99-11e9-a5b3-34f3edf1351e_story.html.

122. Danny Vinik, “What happened to Trump’s war on data?” *Político*, July 25, 2017, https://www.politico.com/agenda/story/2017/07/25/what-happened-trump-war-data-000481/.

123. “A Vision for Equitable Data: Recommendations from the Equitable Data Working Group” (Findings of Working Group, April 2022), https://www.whitehouse.gov/wp-content/uploads/2022/04/eo13985-vision-for-equitable-data.pdf; Daniel Castro, “Biden Creates Road Map for Equitable State and Local Data,” *Government Technology*, March 2021, https://www.govtech.com/opinion/biden-creates-road-map-for-equitable-state-and-local-data.html.

124. Swapnil Kangralkar, “Types of Biases in Data,” *Towards Data Science,* August 26, 2021, https://towardsdatascience.com/types-of-biases-in-data-cafc4f2634fb.

125. “How can you statistically correct for missing data and selection bias in HIV prevalence estimates?” Harvard Center for Population and Development Studies, last modified February 2015, https://www.hsph.harvard.edu/population-development/tag/selection-bias/; Apoorva Mandavilli, “Half of H.I.V. Patients Are Women. Most Research Subjects Are Men,” *The New York Times*, May 28, 2019, https://www.nytimes.com/2019/05/28/health/women-hiv-trials.html.

126. Bella Struminskaya et al., “Sharing Data Collected with Smartphone Sensors: Willingness, Participation, and Nonparticipation Bias,” *Public Opinion Quarterly,* 85:1 (2021), 423–462,

https://academic.oup.com/poq/article/85/S1/423/6363382?login=true .

127. Catherine D'Ignazio and Lauren Klein, "What Gets Counted Counts" in *Data Feminism* (MIT Press, 2020), https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp/release/3.

128. Darcell Scharff et al., "More than Tuskegee: Understanding Mistrust about Research Participation" (Journal of Health Care for the Poor and Underserved), 21:3 (2010), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4354806/.

129. "Leveraging Federal Statistics to Strengthen Evidence-Based Decision-Making," www.whitehouse.gov/wp-content/uploads/2022/03/ap_15_statistics_fy2023.pdf.

130. National Academies of Sciences, Engineering, and Medicine, "Current Challenges and Opportunities in Federal Statistics," in *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy* (National Academies Press, 2017), https://www.ncbi.nlm.nih.gov/books/NBK425871/.

131. Ibid.

132. "Infrastructure Series: Challenges Loom for Chief Statistician of the US," *AmStatNews*, April 1, 2022, https://magazine.amstat.org/blog/2022/04/01/infrastructure-series-csotus/.

133. National Academies of Sciences, Engineering, and Medicine, "Coordination and Collaboration with Other Statistical Agencies," in *Principles and Practices for a Federal Statistical Agency* (National Academies Press, 2021), https://www.ncbi.nlm.nih.gov/books/NBK573395/.

134. "Hard-to-Count Communities in the 2020 Census," Georgetown Law Center on Poverty and Inequality, accessed June 2022, https://www.georgetownpoverty.org/issues/democracy/census-2/hard-to-count/.

135. "Census 2020: How to Count Hard-to-Count Communities," National League of Cities, accessed June 2022, https://www.nlc.org/article/2019/05/03/census-2020-how-to-count-hard-to-count-communities/.

136. Eline Chivot, "What Other Countries Can Learn from the UK's Data Strategy" (Center for Data Innovation, December 2020), https://datainnovation.org/2020/12/what-other-countries-can-learn-from-the-uks-data-strategy/.

137. Hodan Omaar, "Response to the Euroepan Commission's Consultation on the European Strategy for Data" (Center for Data Innovation), https://www2.datainnovation.org/2020-eu-data-strategy.pdf.

138. "How does the federal government define 'disability?'" Department of Labor, accessed June 2022, https://www.dol.gov/agencies/odep/publications/faqs/general#3.

139. "Using Citizen-Generated Data to monitor the SDGs," DataShift, accessed June 2022, https://www.data4sdgs.org/sites/default/files/2017-09/Making%20Use%20of%20Citizen-Generated%20Data%20-%20Data4SDGs%20Toolbox%20Module.pdf.

140. Marina Tavra, Ivan Racetin, and Josip Peros, “The role of crowdsourcing and social media in crisis mapping: a case study of a wildfire reaching Croatian City of Split,” *Geoenvironmental Disasters* (2021), https://geoenvironmental-disasters.springeropen.com/articles/10.1186/s40677-021-00181-3.

141. Amelia Hunt and Doug Specht, “Crowdsourced mapping in crisis zones: collaboration, organisation, and impact,” *Journal of International Humanitarian Action* (2019), https://jhumanitarianaction.springeropen.com/articles/10.1186/s41018-018-0048-1.

142. “Disaster Activation: Nepal Earthquake 2015,” Humanitarian OpenStreetMap Team, accessed June 2022, https://www.hotosm.org/projects/nepal_2015_earthquake_response.

143. “Crowdsourcing Map Data for Humanitarian Response and Preparedness,” CitizenScience.gov, accessed May 2022, https://www.citizenscience.gov/mapgive/#.

144. “Participatory Science for Environmental Protection,” Environmental Protection Agency, accessed June 2022, https://www.epa.gov/citizen-science/; Gillian Diebold, “Citizen Science and Crowdsourced Data Can Improve Environmental Data in the United States” (Center for Data Innovation, June 2022), https://datainnovation.org/2022/06/citizen-science-and-crowdsourced-data-can-improve-environmental-data-in-the-united-states/.

145. Stefan Jungcurt, “Citizen-Generated Data: Data by people, for people,” IISD, May 24, 2022, https://www.iisd.org/articles/insight/citizen-generated-data-people.

146. John Stevens, “The Potential for Alternative Data in Official Statistics” (presentation at Federal Economic Statistics Advisory Committee Meeting, June 2020), https://apps.bea.gov/fesac/meetings/2020-06-12/Stevens.pdf.

147. Hodan Omaar and Daniel Castro, “Comments to OSTP and NSF on a National AI Research Resource (NAIRR),” September 28, 2021, https://datainnovation.org/2021/09/comments-to-the-ostp-and-nsf-on-anational-airesearch-resource-nairr/.

148. Tajha Chappellet-Lanier, “AI development requires good datasets, and OMB wants ideas on how to help,” *FedScoop*, July 11, 2019, https://www.fedscoop.com/america-ai-initiative-data-omb-rfi/.

149. “Participatory approaches to transparency in dataset documentation,” Data Cards Playbook, accessed June 2022, https://pair-code.github.io/datacardsplaybook/.

150. “About ImageNet,” ImageNet, last modified March 2021, https://www.image-net.org/about.php.

151. “MIT researchers find ‘systematic’ shortcomings in ImageNet data set,” *Venture Beat*, July 15, 2020, https://venturebeat.com/2020/07/15/mit-researchers-find-systematic-shortcomings-in-imagenet-data-set/.

152. Edwin Chen, “30% of Google’s Emotions Dataset is Mislabeled,” *SurgeAI*, July 11, 2022, https://www.surgehq.ai//blog/30-percent-of-googles-reddit-emotions-dataset-is-mislabeled.

153. Maria Mestre, “Avoiding top pitfalls in annotation projects, “ *Towards Data Science*, February 21, 2022, https://towardsdatascience.com/avoiding-top-pitfalls-in-annotation-projects-a3165c5e278f.

154. “GitHub Issues,” GitHub website, accessed June 2022, https://github.com/features/issues.

155. “Casual Conversations Dataset,” Meta AI, last modified April 2021, https://ai.facebook.com/datasets/casual-conversations-dataset/.

156. Daniel Castro, “The Rise of Data Poverty in America.”

157. “Homework Gap and Connectivity Divide,” Federal Communications Commission, accessed April 2022, https://www.fcc.gov/about-fcc/fcc-initiatives/homework-gap-and-connectivity-divide.

158. Joshua New, Daniel Castro, and Matt Beckwith, “How National Governments Can Help Smart Cities Succeed” (Center for Data Innovation, October 2017), https://www2.datainnovation.org/2017-national-governments-smart-cities.pdf.

159. Colin Cunliff, Ashley Johnson, and Hodan Omaar, “How Congress and the Biden Administration Could Jumpstart Smart Cities With AI” (ITIF, March 2021), https://itif.org/publications/2021/03/01/how-congress-and-biden-administration-could-jumpstart-smart-cities-ai/.

160. Daniel Castro, “The Rise of Data Poverty in America.”

161. “Executive Order on Advancing Racial equity and Support for Underserved Communities Through the Federal Government” (Executive Order, January 20, 2021), https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/.

162. “Federal Data Strategy,” data.gov, accessed July 2022, https://strategy.data.gov/.

163. “The Data Equity Framework,” We All Count, accessed June 2022, https://weallcount.com/the-data-process/.

ABOUT THE AUTHOR

Gillian Diebold is a Policy Analyst at the Center for Data Innovation. She holds a B.A. from the University of Pennsylvania, where she studied Communication and Political Science.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation is the leading global think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C., and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the nonprofit, nonpartisan Information Technology and Innovation Foundation (ITIF).

contact: info@datainnovation.org

datainnovation.org