Atlantic Council
GEOTECH CENTER

THE DATA DIVIDE

How Emerging Technology and its Stakeholders can Influence the Fourth Industrial Revolution

Joseph T. Bonivel Jr. Ph.D
Solomon Wise

MISSION STATEMENT

Shaping the Future of Technology and Data together to advance people, planet, prosperity, and peace.

Atlantic Council
GEOTECH CENTER

THE DATA DIVIDE

How Emerging Technology and its Stakeholders can Influence the Fourth Industrial Revolution

Joseph T. Bonivel Jr. Ph.D
Solomon Wise

ISBN: 978-1-61977-251-9

Cover image: Used by permission

October 2022

TABLE OF CONTENTS

1. THE DIGITAL DIVIDE

Between the late 1960s and mid-2010s, the Third Industrial Revolution brought microprocessors, personal computing, and the Internet that created an ecosystem that exponentially increased communication capabilities throughout the world. Computers went from expensive and hard to use, room-sized mainframe machines to inexpensive personal computers, beginning with the Commodore 64, and leading to handheld personal devices that contain more processing power than all their predecessors combined. In conjunction with computer advances, the design and use of computer technology—known as human-computer interactions—evolved from Herman Hollerith's "punch" cards to keyboards and mice, and currently operates with touchscreen and voice-activated commands. Advances have allowed for exponential increases in humans' ability to communicate with each other, starting with the telegraph and ending with instantaneous communication via text messaging and video chats. APRANET came online in 1969 with speeds of 56 kilobits per second (Kbps) and was used to connect government agencies and universities focused on defense research; currently, companies offer one gigabyte per second (GBps) fiberoptic speeds, allowing for on-the-spot worldwide knowledge access.[1]

The United States' first great step in this computing evolution was the passage of the High-Performance Computing Act (HPCA) of 1991.[2] HPCA ushered in the necessity of a National Information Infrastructure and provided the funding for the National Research and Education Network (NREN), which focused on providing access to the Internet for all K–12 students. NREN provided a collaboration tool that teachers utilized to share pedagogical tools and methodologies. From 1994 to 1997, the National Science Foundation and the HPCA funded the development of the high-speed research network that would eventually become the Internet.[3] During this same time period (1991–1996), the number of personal computers in the United States increased from three hundred thousand to more than ten million.[4] By the mid-1990s, the development of Internet browsers enabled computers and information to be transmitted via a new realm: cyberspace.[5] The ability to transmit information at high speeds led to the development of a message-delivery system dubbed email, which became increasingly useful due to its speed and widespread accessibility. During this Internet boom, the Bill Clinton administration began to investigate whether access to information technology was being evenly distributed throughout society.[6]

In 1995, the new National Telecommunications and Information Administration (NTIA) produced a report, "Falling Through the Net: A Survey of the 'Have Nots' in Rural and Urban America," which focused on the penetration and usage of information and communication technology (ICT) throughout the United States. The report found that people who did not have access to ICT were disproportionately based in rural areas, and education was correlated with access to the telephone, computer, and household computer modem.[7] The NTIA report is regarded as one of the first instances in which the federal government recognized policies were needed to curb inequalities in access to the Internet, and it was one of the first descriptions of what is now known as the digital divide, or a disparity in the access to, use of, or impact of ICT.

In 1995, whites owned computers at three times the rate of African Americans and Latinos.[8] As Internet connections became more common, demographic groups including African Americans, Latinos, non-English-speaking Asians, tribal and rural populations, the elderly, and adults living with disabilities were slower to adopt the technology. This delay effectively locked them out of the Internet "boom" and its corresponding opportunities for growth and advancement.[9] In 2010, the Pew Research Center found that

1 Randolph A. Miller and Edward H. Shortliffe, "Donald A.B. Lindberg and the U.S. National Library of Medicine Transformed Biomedical and Health Informatics," Information Services & Use 42, 1, May 10, 2022, 3–10, *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9116201/*.

2 Donald A. B. Lindberg and Betsy L. Humphreys, "The High Performance Computing and Communications Program, the National Information Infrastructure, and Health Care," Journal of the American Medical Informatics Association 2, 3 (1995), *https://academic.oup.com/jamia/article-abstract/2/3/156/875750?redirectedFrom=fulltext*.

3 K. M. Hayes and R. J. W. Cline, "Consumer Health Information Seeking on the Internet: the State of the Art," Health Education Research 16, 6 (2001), 671–692, *https://pubmed.ncbi.nlm.nih.gov/11780707/*.

4 Aaron Weiss, "Computing in the Clouds," NetWorker 11, 4 (2007), 16–25, *https://dl.acm.org/doi/fullHtml/10.1145/1327512.1327513*.

5 Ibid.

6 Ibid.

7 Ronald H. Brown, David J. Barram, and Larry Irving, "Falling Through the Net: A Survey of the 'Have Nots' in Rural and Urban America," US Department of Commerce, 1995, *https://www.ntia.doc.gov/ntiahome/fallingthru.html*.

8 Miller, "Donald A.B. Lindberg and the U.S. National Library of Medicine Transformed Biomedical and Health Informatics."

9 Robert Branson, Danielle Davis, and Marcella Gadson, "Bridging the Digital Divide," Multicultural Media, Telcom and Internet Council, 2022, *https://www.benton.org/headlines/wireless-communities-color-bridging-digital-divide*.

the top reasons households could not, or chose not to, get access to the Internet were

- They did not have service available in their area;
- They could not afford it;
- They did not understand how to use it;
- They did not trust it; and
- They did not see its usefulness.[10]

Together, these factors created, and later deepened, the digital divide for many of these marginalized communities, and continued to increase the wealth gap and socioeconomic status disparities.

Fortunately, technology is much more widespread today than it was in 1995. Accessibility has increased, as innovation brought reductions in the costs of central processing unit (CPU) memory, storage, and processing power. Modern ICT, particularly mobile devices with wireless connectivity, has been championed as a bridge across the digital divide. Today, more than 91 percent of adults are connected via wired or wireless broadband, and 85 percent have a smartphone, with more than 20 percent using smartphones solely for broadband Internet access.[11] The high rate of adoption of smartphones and their connection to wireless broadband have granted Internet access to more than three hundred and fifteen million people across the country, and helped to narrow the digital divide. Increased adoption of wireless connectivity by minority groups, dubbed by the Cellular Telecommunications and Internet Association as "The Minority Wireless Miracle," is due to wireless' innate mobility-based flexibility, varied pricing tiers, and widespread coverage.[12] People of color have over-indexed wireless Internet usage since tracking began by Pew in 2011; African Americans and English-speaking Latinos are among the most active users of the mobile Internet. In addition, compared to white populations, members of these groups are more likely to own a cellphone but no personal computer (PC).[13] In many cases, cellular access is the only lifeline to the Internet for disenfranchised groups, allowing them to be part of the digital ecosystem.

2. THE GENESIS OF THE DATA DIVIDE

In 1965, Gordon Moore predicted that the number of components in an integrated circuit would double each year for the next ten years—and reach an astonishing sixty-five thousand parts by 1975.[14] Moore's prediction was validated in 1975 and became the "golden rule" in chip manufacturing, becoming known as Moore's Law. Moore's Law states that the number of transistors on a microchip will double every two years, and that exponential growth in microprocessors will thereby increase computing power. Since then, his prediction has defined the trajectory of technology, ushered in the Third Industrial Revolution (characterized by electronics and information technologies), and introduced the Fourth Industrial Revolution.[15]

The Fourth Industrial Revolution, which began in 2016, is an integration of the cyber-physical world. It is an amalgamation of technologies that focus on physical, digital, and biological spheres. Access to low-cost, low-power sensors, standards for accessing the Internet, cloud-computing platforms, machine learning (ML), and artificial intelligence (AI) have enabled the creation of ICT that touches every segment of daily life. The Internet of Things (IoT), billions of low-cost sensors and people connected by mobile devices (or Internet), is currently producing 2.5 quintillion bytes of "big data" daily. Big data differ from the data in the Third Industrial Revolution in their volume, speed of creation, and dissemination, and their variety creates endless opportunities for process inputs to emerging technologies. Big data—along with unrivaled computer processing power, limitless storage capacity, and instant access to knowledge—form the foundations for advanced prediction algorithms. Emerging technologies such as ML, AI, advanced manufacturing, IoT, nanotechnology, biotechnology, energy storage, and quantum computing have the capability to advance global prosperity and development. These breakthroughs have transformed entire systems of production, healthcare, and governance.[16]

This Fourth Industrial Revolution offers an unprecedented opportunity not only to improve the quality of life, but to close societal gaps. People with access to the digital world

10 Aaron Smith, "Home Broadband Adoption 2010," Pew Research Center, August 11, 2010, *https://www.cetfund.org/report/2010-home-broadband-adoption/*.

11 Andrew Perrin, "Mobile Technology and Home Broadband 2021," Pew Research Center, June 3, 2021, *https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/*.

12 Branson, et al., "Bridging the Digital Divide."

13 Smith, "Home Broadband Adoption 2010."

14 Gordon E. Moore, "Cramming More Components onto Intergrated Circuits," Electronics, 38, 8 (1965), *https://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf*.

15 David Rotman, "We're Not Prepared for the End of Moore's Law," MIT Technology Review, February 24, 2020, *https://www.technologyreview.com/2020/02/24/905789/were-not-prepared-for-the-end-of-moores-law/*.

16 Klaus Schwab, "The Fourth Industrial Revolution: What It Means, How to Respond," World Economic Forum, January 14, 2016, *https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/*.

have benefited the most as technology has created products and services that improve our lives. High-speed mobile technologies and wireless services allow us to connect with services for transportation, groceries, products, entertainment, health information, and personal connections. The IoT, a collection of Internet-connected devices and cloud services, has made it possible to gather information, analyze it, and use it to act. The collection of big data from mobile applications, Internet browsing, and IoT devices, in conjunction with machine learning and artificial intelligence, has made it possible to build a predictive computational theory of human behavior that can be used for the benefit of humanity or for nefarious reasons. Applications that utilize AI are often deployed for expeditious decision support and decision-making, to remove the human from the loop.[17]

Machine learning is a method of data analysis in which non-human systems gather insights from data and recognize patterns, and is considered the basis of artificial intelligence.

Artificial intelligence refers to a software-based system that receives and makes decisions based on signals from the external environment. AI also takes actions that affect the external environment by generating outputs such as predictions, recommendations, or decisions based on incoming data.[18]

Data are the new oil of the digital economy—they are crucial to the global economic framework, impacting everything from credit rates, consumer-marketing tactics, and judicial sentencing guidelines to local, state, and federal elections. Those entities that have functional access to data capital have more options than those that do not. The data divide is the gap that exists between individuals who have access, agency, and control with respect to data—and can reap the most benefits from data-driven technologies—and those who do not.

The data divide is a secondary effect of the digital divide, and manifests itself in the way that data systems are designed and developed. Those who have access to these data systems are most likely to be represented in the outputs of the machine-learning and artificial-intelligence systems. Within the United States, the data divide most impacts the underserved, underprivileged communities that lack adequate resources to access data, or are otherwise prevented from making constructive use of data and the data decision-making process. The data divide has a determining effect on who can be represented by and can shape data-driven technologies, which perpetuates and compounds social and health inequalities.

According to surveys conducted during the COVID-19 pandemic by the Ada Lovelace Institute, key elements of the data divide's widening gap include the following.

- Differential access: people without fundamental access to technologies that bring them online are invisible to data processes.
- Differential knowledge, awareness, and skills: people may not be aware of the tools available, based on education or reduced digital literacy.
- Differential trust: even if they have access to, and knowledge of, data-driven technologies, there may be historic and structural reason underrepresented groups choose not to be involved in the data-driven society.[19]
- Unlike the digital divide, where lack of access to the Internet can be attributed to social, economic, and geographic factors that can be remedied, the data divide has the risk of widening to a point at which it will not be possible to close it.

The emerging data divide exists across multiple dimensions, and is affected by multiple stakeholders. Internationally, data capabilities are primarily developed in the Global North (with China a notable exception), as those countries' governments institute policies and programs to reap social impact and critical-infrastructure benefits from big data/ML/AI. Governments collect substantial amounts of data on their citizens, including name, age, voting record, birthdate, occupation, home location, criminal record, and ethnicity. These data can be used for socioeconomic analysis to benefit their constituents, but can also be used to further disenfranchise selected groups from government benefits and democratic processes. This dimension often maps onto prior inequalities, so that individuals and communities with sparse access to healthcare, civil protections, educational opportunities, or income are also excluded from the increasing benefits of datafication.

Even within countries that are broadly benefitting from big data, a second dimension of the data divide exists between those who have digital literacy and those who do not understand the benefits or exploitation arising from collection and use of their data. Nongovernmental organizations (NGOs), universities, and advocacy groups have been at the forefront of voicing concerns about digital-literacy capacity

17 Reva Schwartz, et al., "Towards a Standard for Identifying and Managing Bias in Artifical Intelligence," National Institute for Standards and Technology, March 2022, *https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf*.

18 Peter Norvig and Stuart J. Russell, Artificial Intelligence: A Modern Approach (Hoboken, NJ: Prentice Hall, 1995).

19 "The Data Divide," Ada Lovelace Institue, March 25, 2021, *https://www.adalovelaceinstitute.org/report/the-data-divide/*.

building and innovations in bias reduction in ML/AI, and highlighting data stewardship in the data lifecycle.

A third dimension exists, with the commercial sector utilizing data for profit from the digital economy and the use of data for noncommercial objectives. The accompanying deployment of big-data technology by businesses has expanded the commercial sector's influence on basic determinants of community well-being, such as healthcare, housing, transportation, education, and food. This expanded influence has not been counterbalanced by actors with noncommercial objectives, because these entities lag in data education, tools, and talent. Reducing the imminent data divide requires understanding of the data lifecycle, data-processing techniques in machine learning and artificial intelligence, and optimization of talent, resources, and tools from stakeholders in the commercial, government, and nonprofit sectors.

2.1 Data Lifecycle

As individuals, communities, and nation-states interact with a world that is becoming increasingly virtual due to the IoT and cellular accessibility, they become vulnerable to the commodification of their digital footprints.[20] Natural tendencies that were previously untracked are easily captured, quantified, and modeled to predict human behavior. These models determine whether someone will get a mortgage loan or acceptance into a prestigious university, or develop health conditions that may keep them from employment. While many organizations seek to use data for good and practice data stewardship, biases remain common across technological processes, and can result in detrimental impacts—whether intentional or not.[21] The data used in these predictions hold the key to determining socioeconomic mobility for many. As a result, bridging the data divide requires that government, industry, and academics have in-depth knowledge of the data-lifecycle process.

In general, the data-lifecycle process is an approach to managing data from entry to destruction. Data are separated into phases based on a set of criteria, and move through these stages as they meet new requirements or complete tasks. This process has been optimized for commercial value, and stakeholders covet their individual data-lifecycle processes as intellectual property. The reduction of the data divide will require optimization of this process, specifically for social issues.

For this report, the data-lifecycle process has been simplified to four stages highlighted in Figure 1. This outline of the data process, while not exhaustive, outlines the major things to consider when utilizing data for ML/AI.

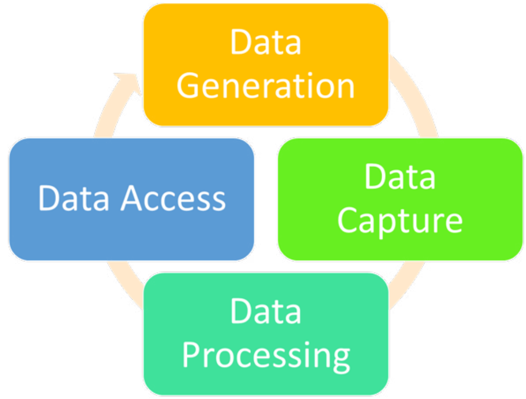

Figure 1. Data Lifecycle

> *Bias: Bias in data collection causes datasets to be statistically unrepresentative, and not generalizable to wider populations.[22]*

Typical data-lifecycle management processes incorporate data maintenance, storage, use, and archival or destruction. These steps, along with intermediary analysis and cleaning included in data processing, are summarized below, with an eye toward the data divide.

2.1.1 Data Generation

Data generation is the first step in the data lifecycle, and the beginning of big data. Such generation is easy to come by—generation occurs regardless of whether the user is aware of it, through every online sale, purchase, hire, communication, or social media interaction. These data can bring about powerful insights that can be utilized by ML/AI and allow for commercial and socioeconomic impacts.

> *Projections indicate that by the end of 2025, the world will potentially generate one hundred and eighty-one zettabytes (one hundred and eighty-one trillion gigabytes) of data, an increase of more than 100 percent from the seventy-nine zettabytes generated by 2021.*

Data Generation Concerns

A sizable portion of the data divide originates before data capture. Legacy data are historical data that were generated before the Third or Fourth Industrial Revolution and have both positive and negative effects on the data divide.

20 Schwartz, et al., "Towards a Standard for Identifying and Managing Bias in Artifical Intelligence."

21 Ibid.

22 "Glossary of Statistical Terms," Organisation for Economic Co-operation and Development, July 2007, *https://stats.oecd.org/glossary/detail.asp?ID=3605*.

Legacy data can date back centuries when recordings were manually generated and saved (using human-centric, labor-intensive inputs). Legacy data are the backbone of ML/AI and automation—and if legacy data are not combined with real-time data, it will be impossible to get the right context and value. Legacy data can provide ML/AI models with inputs that help determine climate baselines or trace epidemics in specific regions and times in history (such as the bubonic plague) that can inform pandemic response in the future. However, the challenges with legacy data include both the volume and representation in the data. Negative legacy data can intentionally or unintentionally contain bias that alienates a group, region, or gender. This bias can be propagated as training data for an ML algorithm, and can generate an incorrect or biased result.

In a case study conducted by Princeton University, ML models trained using biased legacy data performed language-translation operations that associated female names with characteristics such as "parents" and "weddings," while male names had stronger correlations with words such as "professional" and "salary." The model picked up this correlation based on legacy data mined from text that reflected these gender tropes.[23] Within natural language processing, a subfield of human-computer interaction (HCI), AI, and linguistics, gender bias is a concerning but well-researched challenge, and understanding the legacy-data inputs that lead to this gender bias provide the path to correct it. In non-gendered languages, such as English, researchers have found methods to enforce word embeddings that are gender neutral.[24] In cases where language is inherently gendered, corpus linguistics can be used to prevent bias by introducing new counter-examples that break causal relationships between gendered and gender-neutral words.[25]

The IoT and ICT provide a plethora of avenues for data generation for ML/AI techniques to utilize. These avenues must be combined with quality legacy data to take full advantage of ML/AI and create reliable and unbiased algorithms. To ensure the quality of legacy data, data scientists must identify which data are not represented in datasets.

2.1.2 Data Capture

In our current digital age, data generation can happen autonomously, but not all data that are generated are collected or used. Data capture, or the selection of data for usage, is determined by the specifications and model by which the data will be analyzed. This requires an in-depth understanding of what commercial, governmental, or societal challenge is to be addressed. Once these requirements are set, the best means of capturing the data can be standardized so they can be accessible and manageable at later stages. The purpose of data capture is to translate information from all sources into a format that computers can understand, without the inclusion of redundant or unnecessary details.

Data from mobile technologies—including purchases, location, social media activity, and even keystrokes—can be captured and automatically formatted for commercial use. There are several data-capture methods for documenting information from traditional data-capture methods, including surveys, emails, invoices, and other sources. These methods are classified into two types: manual and automatic. The manual data-capture process is an antiquated, labor-intensive, human-in-the-loop technique of obtaining and manually inputting information utilizing media such as pen, paper, keyboards, and touch displays. Automated data capture utilizes advanced technologies, such as optical character recognition, bar codes, digital signatures, and intelligent document processing.[26]

It is important to note that many organizations take a broad approach to the collection of data, gathering the maximum possible amount of data from each data-generating interaction and storing all of them for potential future use. Though drawing from this broad supply is useful for ML/AI, it is most efficient to create a plan to capture all the data needed for specific analyses, such as those related to climate change, financial reporting, and health outcomes.

Data Capture Concerns

IoT devices, particularly sensors, are now continuously producing data. Whether from a smart thermometer or smart watch—and regardless of whether the IoT platform is a drone (data generation) or edge (data collection and processing) device—everything can generate larger amounts of data. Without cloud computing, the data-management industry was unable to capture this data generation, either through networks, fifth-generation (5G) technology, cloud

23 Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," Science, April 14, 2017, 183–186, *https://www.science.org/doi/10.1126/science.aal4230*.

24 Anupam Datta, "3 Kinds of Bias in AI Models—and How We Can Address Them," InfoWorld, February 24, 2021, *https://www.infoworld.com/article/3607748/3-kinds-of-bias-in-ai-models-and-how-we-can-address-them.html*.

25 Ibid.

26 Haissam Abdul Malak, "What Is Data Capture and Why Is It Important?" ECM Consultant, October 7, 2021, *https://theecmconsultant.com/what-is-data-capture/*.

computing, or any other storage method.[27] These technical challenges led to 90 percent of captured data being lost due to an inability to store and rapidly process them, which revisits the problem of data bias and data provenance.

As a result of data being removed or lost, the data that are collected and possibly implemented as training data may be significantly different from the full dataset that was collected.[28] Sampling bias occurs when data are intentionally or accidently removed, resulting in certain members of a population being more likely to be selected in a sample than others. The training datasets are based on samples that are neither properly randomized nor truly representative of the population sampled.[29] Datasets suffering from sampling bias are not generalizable, yet are often used to train ML/AI applications that are deployed for use in socioeconomic predicative contexts, such as creditworthiness or healthcare outcomes, despite the exclusion of data representing certain populations.[30]

Understanding how data were captured allows data scientists to understand the entire ecosystem around them. Data provenance refers to the documentation of where a piece of data originates, and the methods by which it was produced.[31] Recording data provenance is necessary to confirm the authenticity of a dataset and to enable it to be reused.

Data provenance history: Provenance, as a practice, has been used in the context of art history to document the history of an artwork, and in digital libraries to document a digital object's lifecycle.

For ML/AI training and modeling, it is always better to have too much data than too little. When more data are presented, the algorithm will create more connections between neurons. Eliminating too much generated data from data collection generates algorithmic bias, which propagates through outcomes.

2.1.3 Data Processing

After data are captured, they must be processed in a number of ways.

- Data wrangling: a dataset is transformed from its raw form into a more broadly accessible format.
- Data compression: data are transformed into a format that can be stored more efficiently.
- Data encryption: data are transformed into another form of code to protect them from unauthorized access. This process is paramount in addressing privacy concerns.[32]

While this list is not comprehensive of all the steps in the data-processing domain, this report focuses on data-processing concerns in the context of bridging the data divide. Although standardization during data capture is prescribed, ICT sources typically require data preparation when combining multiple data sources (e.g., location, purchase habits, social media posts) to make predictive models. The data-preparation process, which must occur prior to processing, is one of the main challenges for data scientists.

According to a recent study, data preparation (i.e., putting data into usable formats) is the most labor-intensive portion of time spent on ML initiatives. Data scientists spend most of their time on data cleaning (25 percent), labeling (25 percent), augmentation (15 percent), aggregation (15 percent), and identification (5 percent).[33] Lack of necessary data, data that are not ready for use, incompatible data formats, unstructured data, and unbalanced data are major challenges that affect data processing and make it time consuming.

Data Processing Concerns

While government and corporate policies can be

27 Sahil Chawla, "Artificial Intelligence: The Future Is Data Capture, Not Machine Learning," Times of India, February 21, 2022, *https://timesofindia.indiatimes.com/blogs/voices/artificial-intelligence-the-future-is-data-capture-not-machine-learning/*.

28 Abigail Z. Jacobs and Hanna Wallach, "Measurement and Fairness" in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, *https://dl.acm.org/doi/10.1145/3442188.3445901*.

29 Ibid.

30 Catherine D'Ignazio and Lauren F. Klein, Data Feminism (Cambridge, MA: MIT Press, 2020), *https://data-feminism.mitpress.mit.edu/*.

31 Ashley Hay, "What Is Data Provenance?" About Data Provenance, last visited August 30, 2022, *http://faculty.washington.edu/hazeline/ProvEco/generic.html*.

32 Tim Stobierski, "Data Wrangling: What It Is and Why It's Important," Harvard Business School, January 19, 2021, *https://online.hbs.edu/blog/post/data-wrangling*.

33 Alfrick Opidi, "Solving Data Challenge in Machine Learning with Automated Tools," TOPBOTS, September 19, 2019, *https://www.topbots.com/data-preparation-for-machine-learning/*.

implemented to address concerns related to data generation, capture, and access, most data-processing concerns are tackled via statistical/mathematical methods to avoid model bias. In data processing, the scientific methodologies used to gather, process, and authenticate input data to create an output data product rely on the credibility and trustworthiness of the input data; as a result, checking data quality and precision provides confidence in outputs.

As discussed in section 2.1.2, data capture (or lack thereof) cascades into the data-processing concerns; if the data captured are not an adequate representation of the population, the outcome will have bias. While not the focus of this report, a common method for addressing AI bias is to focus on creating equitable statistical representation in the legacy datasets used in AI training processes. Techniques such as class-imbalance measures, label-imbalance measures, or analysis using the statistical Simpson's Paradox can be used to detect and mitigate bias in datasets that are used to train AI.[34]

2.1.4 Data Access

Data processing and data access are a continuous feedback loop through which data become available to users. Data-access gatekeepers must define who can use the data and the purpose for which they can be used.[35] Once data are made available, they can be leveraged for a variety of analyses, ranging from basic exploratory data processes and data visualizations to more advanced data mining and ML/AI techniques.

Data Access Concerns

Of all the concerns within the data lifecycle related to the data divide, data access is the most challenging. Its direct-feedback correlation to data processing, in conjunction with model predication outcomes, provides the greatest opportunity for impacting the data divide. The potential misuse of personally identifiable information (PII) in decision-making creates mistrust from disenfranchised/underrepresented groups that have historically had their data shared for purposes that did not benefit the group. According to a Pew Research Center survey, a statistically significant portion of respondents reported being concerned about how their data are being used by private-sector firms (79 percent) or the government (64 percent).[36] Respondents also expressed concern about their lack of control over the use of their personal information by these entities.[37]

3. MAJOR STAKEHOLDERS IN THE DATA DIVIDE AND THEIR INFLUENCE

A discussion of the data divide would not be complete without an overview of the major stakeholders in the data sphere and their influence. This section will discuss three key stakeholders with respect to the data divide: private-sector corporations, governments, and civil-society organizations such as NGOs, advocacy organizations, and academic institutions. For each stakeholder group, this section will present a model of behavior as relates to the data divide, including organizational priorities, potential room for improvement, and opportunities and policies that should be pursued to assist or incentivize each stakeholder in bridging the data divide.

3.1 - Private Sector

The purpose of the firm has long been debated among practitioners and scholars alike. Proponents of shareholder theory argue that the sole objective of the firm is to maximize its value for shareholders.[38] This objective takes precedence over the interests of other stakeholders in the firm's practice, such as its employees, customers, and society at large. Others argue that a firm should seek to satisfy the interests of all groups with a stake in the firm, in addition to its shareholders (such as its employees, customers, and suppliers), to ensure the long-term success of the firm.[39] Others include yet more stakeholders, claiming that firms should consider their ethical and social obligations to the society in which they operate, and utilize parts of their business for achieving social good.[40]

34 Schwartz, et al., "Towards a Standard for Identifying and Managing Bias in Artifical Intelligence."

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

36 Ibid.

37 Ibid.

38 Manuel Castelo Branco and Lucia Lima Rodrigues, "Positioning Stakeholder Theory Within the Debate on Corporate Social Responsibility," Electronic Journal of Business Ethics and Organization Studies 12, 1 (2007), *https://psycnet.apa.org/record/2007-15413-001*.

39 R. Edward Freeman and John F. McVea, "A Stakeholder Approach to Strategic Management," Darden Graduate School of Business Administration, University of Virginia, July 5, 2022, *https://www.researchgate.net/publication/228320877_A_Stakeholder_Approach_to_Strategic_Management*.

40 Archie B. Carroll, "Carroll's Pyramid of CSR: Taking Another Look," International Journal of Corporate Social Responsibility, 2016, *https://jcsr.springeropen.com/articles/10.1186/s40991-016-0004-6*.

The shareholder theory of firm behavior has serious and worrying implications for the role of private-sector corporations in producing and bridging the data divide. It implies that private-sector corporations have no responsibility to society as relates to the unequal distribution of access, agency, and control over data, and, thus, would not take action to reduce the data divide unless these actions were profitable for the firm. Furthermore, it is possible that, if this theory of the firm holds true, firms may even capitalize on the data divide in their pursuit of maximum shareholder value. This behavior has already been observed in the financial sector, where high-frequency trading firms take advantage of disparities in data access among firms, as well as between firms and consumers. High-frequency trading firms invest hundreds of millions of dollars in infrastructure to receive data on market activity quicker than rival high-frequency trading firms and traditional banks that manage the money of consumers. The use of multiple trading venues (in the United States alone, there are sixteen different stock exchanges and more than fifty alternative trading venues) allows for "latency arbitrage," a process whereby, when the price of a stock shifts in one venue, high-frequency trading firms race to either buy from the venue and sell on others if the price has decreased, or sell on the venue if the price has increased, while consumers and traditional banks are none the wiser.[41] This multibillion-dollar industry depends on inequality in access to, and control over, financial data among market participants.

However, as previously noted, the view that the sole objective of the firm should be to maximize shareholder value is disputed both by scholars and by the business community. Throughout history, firms have been observed behaving in ways that benefit not only their shareholders, but society more broadly. In this practice, known as stakeholder capitalism, firms prioritize not only short-term profits for shareholders, but long-term value creation for the whole of society, by considering the needs of a broader range of stakeholders including employees, suppliers, customers, the state, and civil society.

Stakeholder capitalism holds the key to unlocking the power of the private sector in bridging the data divide. Although actions by firms that are motivated solely by shareholder-value maximization may incidentally play a role in bridging the gap between those who have access, agency, and control over data, more purposeful and targeted actions are needed.

Data-sharing initiatives are an excellent way for private-sector firms to bridge the data divide. Private-sector firms capture tremendous amounts of customer data, which can be useful not only for the corporation's bottom-line profits, but also for society. A notable success in for-profit private-sector data sharing took place in the aftermath of Hurricane Katrina. Valassis, a direct-mail marketing company, shared its address database—normally used to send advertisements through the mail—with the nonprofit Greater New Orleans Community Data Center for a nominal fee, allowing it to track the city's recovery process. Valassis' dataset allowed emergency workers to identify households that had returned to the city in the aftermath of the hurricane according to whether they were receiving Valassis' promotional mailings; this permitted the redirection of staff and funds that would have been used for street-by-street, in-person repopulation surveys toward rebuilding efforts. Valassis also shared its data with another nonprofit, Kingsley House, which used the data to identify repopulated areas of the city and enroll children in health-insurance programs.[42] This data-sharing example also sheds light on disparities of access, agency, and control over data between private-sector firms and governments. Valassis had significantly more detailed and updated data on New Orleans addresses and their residents than did the city of New Orleans itself. And although the US Postal Service had a similarly granular dataset, legislation prevented it from sharing that dataset with nonprofits.

3.2 - Government

Government plays a dual role in the data divide. Governments act as the sole arbiters of legislation around data, and have tremendous power to shape the distribution of access, agency, and control over data in society. Simultaneously, governments collect massive amounts of data that could potentially be used for social good, and do not need to consider profit and shareholder value when deciding whether to share data. Thus, governments can act in two ways to reduce the data divide: by using legislation to incentivize firms to share data of social value, and by making more government-collected data publicly available.

Mandatory data-sharing requirements instituted by governments can be a valuable and necessary tool to incentivize private-sector firms to share data that can be used for social good. Take, for example, the utility of tourism data. Economic data around tourism-based service work are not often captured by local, state, and national governments, due to the often-informal nature of the tourism sector. However, these data are valuable to governments in highly tourism-dependent economies, due to tourism's integral role in affecting gross domestic product (GDP) and

41 Matteo Aquilina, Eric Budish and Peter O'Neill, "Quantifying the High-Frequency Trading 'Arms Race,'" Quarterly Journal of Economics 137, 1 (2021), *https://academic.oup.com/qje/article/137/1/493/6368348*.

42 "Valassis Lists Case Studies," Valassis, last visited August 29, 2022, *https://www.valassislists.com/case_study/31*.

impacts on local housing affordability. Digital innovations have made data around tourism-based service work easy to capture by digital services such as Airbnb. In 2017, after extensive negotiations, the city of Portland and Airbnb signed a data-sharing agreement requiring Airbnb to provide regulators with data on rental listings, allowing the city to better monitor the state of the tourism-based economy and crack down on unpermitted rentals that drive up housing prices.[43] This data-sharing agreement would not have been possible if not for legal action taken by the city banning unpermitted rentals, and a subpoena against Airbnb requiring the company to turn over data that the city could use to enforce this legislation.[44]

Making valuable government-collected data widely and publicly available is another way that governments can contribute to bridging the data divide by making the distribution of access and control over data less unequal across society. Positive examples abound of government-collected data being made openly available and subsequently being used for social good. For example, the US Department of Agriculture's Rural Housing Service releases data around the loans, grants, and guarantees the agency gives to nonprofits, state and local agencies, and communities to improve housing and living standards in rural areas, which allows nonprofits and other agencies working on rural housing issues to make more informed, evidence-based decisions.

3.3 - Civil-Society Organizations (CSOs)

NGOs, advocacy organizations, and academic institutions drive policy recommendations to bridge the data divide in their capacity as thought leaders, innovators, and constituent-minded experts. Civil-society organizations are not driven by political agenda or term limits, nor are they focused on profitability and shareholder value. They are uniquely situated to invest in long-term projects that focus on the impact of data and reducing the data divide.

Academic Institutions

The greatest strength of academic institutions lies in their capacity-building capabilities. According to QuantHub, data science ranks among the top technology-related areas in which employers are having difficulty finding enough employees with the appropriate skillsets.[45] This drought is fueled by companies ramping up their data efforts to make sense of newly digitized data, IoT data generation, and cybersecurity concerns, along with small businesses and government agencies seeing the potential of data analytics. Academic institutions are the primary avenue to promote data literacy and bring a new pool of aspiring and diverse junior-level talent into the ML/AI market.

Research plays a meaningful role in finding solutions to ML/AI training-data bias. Academic institutions lead in identifying mathematical solutions and disseminating those solutions through peer-reviewed publications that lead to statistical-methodology adoption. Ensuring datasets are representative and maintaining data provenance should be the primary goal for these bias solutions.

Academic institutions also play a pivotal role (in conjunction with government) in developing data governance for data-processing procedures, thereby ensuring interoperability and scaling for access. For example, the University Corporation for Atmospheric Research, a nonprofit group consisting of more than one hundred and twenty academic institutions focused on research in Earth-systems science, hosts the Network Common Data Form (NETCDF), a set of software libraries and data formats that support the creation, access, and sharing of scientific data. NETCDF has set a community standard for data sharing in the sciences, which enables data under this standard to be shared more easily and on a wider scale. Data adhering to NETCDF include a description of the data they contain, can be accessed by systems with different ways of storing both numbers and text, can be broken down into smaller subsets and accessed via remote servers, and can be added to without copying the dataset or changing its internal structure.[46]

Nongovernmental Organizations

NGOs, especially mission-driven or service organizations, can lead a variety of policy initiatives to diminish the data divide. For example, they may

- lead campaigns at the local, state, and national levels to reduce the data divide;
- promote equitable data-governance policies that ensure diversity and inclusion in datasets in which ML/AI models may have a negative impact on disenfranchised groups;

43 Dan Wu, et al., "How Data Governance Technologies Can Democratize Data Sharing for Community Well-Being," Cambridge University Press, July 13, 2021, *https://www.cambridge.org/core/journals/data-and-policy/article/how-data-governance-technologies-can-democratize-data-sharing-for-community-wellbeing/2BFB848644589873C00E22ADEA6E8AB3*.

44 "Memorandum of Understanding: Pass Through Registration Data Sharing Agreement," City of Portland, August 30, 2019, *http://opb-imgserve-production.s3-website-us-west-2.amazonaws.com/original/airbnb_pass-through_registration_agreement_final_and_signed_1567631972272.pdf*.

45 Jen DeBois, "The Data Scientst Shortage in 2020," QuantHub, April 7, 2020, *https://quanthub.com/data-scientist-shortage-2020/*.

46 "Network Common Data Form (NetCDF)," UCAR Community Programs, last visited August 30, 2022, *https://www.unidata.ucar.edu/software/netcdf/*.

- supply unbiased monitoring and evaluation of programs focused on data for decision-making;
- provide digital-literacy training that allows companies and organizations to become less reliant on small, siloed teams of expensive experts, reducing the data divide by increasing the number of skilled and diverse workers; and
- facilitate data-sharing processes. For example, the Chan Zuckerberg Initiative's Open Science program provides grants to platforms that enable more widespread access and sharing of scientific research data, such as ASAPbio, bioRxiv, and medRxiv, which allow scientists to share important data from their research prior to peer review. These data repositories have played a key role during the COVID-19 pandemic in providing researchers with time-sensitive data more quickly.[47]

Advocacy Groups

While advocacy groups and NGOs share several similarities, and may even have the same objectives, advocacy groups have a special emphasis on altering public policy. Advocacy groups may also work to affect public opinion on data ownership and data transparency by disseminating relevant information about the data divide to constituents in local communities, and ensuring that they become aware and involved in the issue. Most importantly, these groups can directly lobby government leaders to create policies that enforce open data and data stewardship.

For example, the Data Foundation's Data Coalition Initiative brings together leaders from across the data industry to advocate for better data standards and access mechanisms in the US government, particularly data sharing between government agencies, as well as making more government-collected data freely shared and publicly available.[48] The Data Coalition Initiative and similar advocacy groups played a key role in the passage of the Open, Public, Electronic, and Necessary (OPEN) Government Data Act, which codified an "open by default" policy for all government data, and requires federal agencies to release government data in machine-readable and accessible formats.[49]

4. KEY TAKEAWAYS

The Fourth Industrial Revolution is highlighted by the interconnection of devices and sensors to the Internet. The computing and communication capabilities of these devices allow for 2.5 quintillion bytes of data to be produced, stored, and analyzed daily. Approximately 30 percent of the world's data being generated are generated by the healthcare industry. This generation comes from advances in medical equipment, but also from everyday consumer purchases such as smartphones and watches. These data are used as input data into machine-learning and artificial-intelligence models that have strong impacts on multiple healthcare domains that have the potential to impact the socioeconomic statuses of billions of people across the world. Those entities that have the functional access to data capital have more options than those that do not. The data divide is the gap that exists between individuals who have access, agency, and control with respect to data and can reap the most benefits from data-driven technologies, and those who do not. The data divide can only be reduced through optimization in data process, monitoring and evaluation of the policies and programs of major stakeholders, and alignment of public-private partnerships for social good.

Key steps in closing the data divide include

- understanding who or what is not represented in legacy datasets or during data generation to ensure mitigation of bias in ML/AI training datasets;
- recording data provenance that confirms authenticity of the data and builds trust and credibility in the reproducibility of the results from ML/AI training sets;
- requiring balanced statistical representation in the datasets used in modeling processes, to reduce ML/AI statistical bias; and
- ensuring ethical data stewardship for access and privacy concerns in ML/AI-based data operations.

Three stakeholder groups—private-sector firms, governments, and civil-society organizations— have important roles to play vis-à-vis the data divide.

- Private-sector firms capture and process copious amounts of data that are both valuable for their shareholders and socially valuable. When private-sector firms consider the needs of stakeholders aside from their shareholders, these data can be shared with governments and civil-society organizations, and used for social good.
- Governments have a dual role as the sole arbiters of data policy, as well as being major data capturers

47 "Open Science," Chan Zuckerberg Initiative, last visited August 30, 2022, *https://chanzuckerberg.com/science/programs-resources/open-science*.

48 "About Us," Data Coalition, last visited August 30, 2022, *https://www.datacoalition.org/about*.

49 "Open, Public, Electronic and Necessary (OPEN) Government Data Act," 2018, *https://www.congress.gov/bill/115th-congress/house-bill/1770*.

and processors. Policies that incentivize corporations to share socially valuable data, practices that make government-owned data more readily available, and efforts to reduce bias in government-collected datasets are all ways that the government can contribute to bridging the data divide.

- Civil-society organizations of all types have a key role to play regarding the data divide. They can train a new, more inclusive generation of data professionals, create new data-governance structures, and advocate for legislation that will positively affect the distribution of access and control over data across society.

ACKNOWLEDGEMENTS

The GeoTech Center would like to extend its thanks to the interviewees and colleagues whose expert insights and support proved invaluable to this project:

Giulia Neaher, *Assistant Director, Atlantic Council GeoTech Center*

Eleanor Creasey, *Young Global Professional, Atlantic Council GeoTech Center*

Alex Moseson, *Director of Federal Solutions, SFL Scientifid*

Michael Scruggs, *Applied Intelligence Lead, Accenture Federal Services*

Gregory Weiner, *Managing Director, Accenture*

Courtney Lewis, *Program Manager, Amazon Web Services*

DJ Patil, *Former U.S. Chief Data Scientist*

Stuart Gluck, *Advisor to the Director of the Office of Science, US Department of Energy*

Sara Bertran De Lis, *Director of Analytics, Johns Hopkins University Center for Civic Excellence*

Nick Hart, *President, Data Foundation*

Joe Flasher, *Open Data Lead, Amazon Web Services*

Beth Blauer, *Associate Vice Provost for Public Sector Innovation, Johns Hopkins*

Dean Ritz, *Chair of the Board of Directors, Data Foundation*

ABOUT THE AUTHORS

Dr. Joseph T. Bonivel Jr. Ph.D is a Subject Matter Expert for the Department of Defense, where he leads disruptive technologies incubation and maturation for the Undersecretary of Defense for Research and Engineering's Journal of DoD Research & Engineering. He is also a Nonresident Senior Fellow for the GeoTech Center at the Atlantic Council where he provides science policy guidance on emerging technologies and development of strategies to ensure the use of "technology for good" among individuals, societies, and the international community. Joe is a diversity advisory board member for NASA's Translational Research Institute for Space Health, which focuses on solving the challenges of human deep space exploration. He is a former Big Data and Analytics Science and Technology Policy Fellow at the National Science Foundation (NSF) and United States Agency for International Development (USAID). In his NSF fellowship role, he fostered entrepreneurship, innovation, and technology commercialization of research that was previously funded by the federal government. Joe's role at USAID was twofold in helping to understand how digital technology can stop the spread of health epidemics (Ebola) and designing and developing predictive algorithms to measure USAID's influence on their bilateral donors and partners. Prior to his AAAS fellowships, Joe was a Senior Research Engineer for United Technologies Research Center (UTRC). At UTRC he was responsible for developing, identifying, and implementing protocols to evaluate the mechanical performance of novel materials and structures within the various businesses of United Technologies *Corporation.Dr*. Bonivel holds a Ph.D. in Mechanical and Materials Science Engineering from the University of South Florida. He earned a Master of Science degree in Mechanical Engineering at Carnegie Mellon University, and a Master of Science and Bachelor of Science degrees in biomedical engineering and mechanical engineering at the University of South Carolina. Bonivel was recently awarded Technologist of the Year by the Southern New England Association of Technical Professionals and was profiled by Black Enterprise Magazine as one of its 100 Modern Men. Bonivel has also spent time teaching engineering and physics at Federal University of Rio de Janeiro in Brazil.

Solomon Wise is a Program Assistant with the GeoTech Center, where he contributes to projects at the intersection of geopolitics and technology. Prior to joining the Atlantic Council, he pursued a Masters in International Development at the London School of Economics where he researched the impact of emerging technologies on humanitarian governance. In London he also consulted for the Overseas Development Institute and previously interned for the United Nations Democracy Fund in New York. His areas of interest include how emerging technologies are affecting global governance structures, the future of work, and the Global South. Solomon holds a bachelors degree in Political Science from Bucknell University in Pennsylvania.

Atlantic Council Board of Directors

Atlantic Council

The Atlantic Council is a nonpartisan organization that promotes constructive US leadership and engagement in international affairs based on the central role of the Atlantic community in meeting today's global challenges.

Atlantic Council

1030 15th Street, NW, 12th Floor, Washington, DC 20005

(202) 463-7226, *www.AtlanticCouncil.org*