RESEARCH REPORT

Data Philanthropy

Unlocking the Power of Private Data for Public Good

Brice McKeever   *Solomon Greene*   *Graham MacDonald*   *Peter Tatian*
Deondre' Jones
July 2018

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

Contents

Acknowledgments

Introduction

The increasing integration of technology into everyday business practices means that private companies are actively and passively accruing data at a growing rate, creating new opportunities to leverage this wealth of information for the public good. Businesses of all sizes are exploring the frontiers of "data philanthropy" by responsibly sharing their data with researchers, nonprofits, the government, and the public. New approaches to data philanthropy are emerging as data providers experiment with innovative platforms and partnerships. This report discusses some of these approaches, addresses the benefits and challenges associated with data philanthropy, and provides context for exploring new pathways in this emerging field.

The boundaries of data philanthropy have not been fully established, and private companies that provide data often define the practice differently than researchers or beneficiaries.[1] But three common themes cut across most definitions:

1. Private data are generated or collected by a private, for-profit entity, which we refer to as the "data provider" (see box 1 for definitions).

2. The data provider shares data (subject to privacy and data protection safeguards including anonymization and aggregation) or insights generated from its data with the public or a public-serving analyst, such as a research university or nonprofit research institution.

3. The data, or results of analysis performed on the data, are used to yield new insights; improve public policies, programs, and services; or otherwise serve the public good.

By synthesizing these three elements, we define data philanthropy as **the act of sharing private data assets to serve the public good**. This definition allows for considerable latitude in what constitutes data philanthropy. A data provider can be a business or any other private organization that owns, creates, or collects private data. The release of this data can take many forms, ranging from summary reports of internal analyses to opening public access to complete datasets. As the field continues to expand, new partnerships between data providers, data analysts, and data consumers are producing innovative methods for applying privately collected data across a widening array of issues (Future of Privacy Forum 2017). We adopt a broad and flexible definition of data philanthropy to encompass both existing and future applications of private data.

Key Terms

- **Private data:** any data that are not generally available to the public.

- **Data provider:** a private, for-profit entity that creates or collects data (e.g., Twitter, Nielsen).

- **Data analyst:** a person or organization that converts and repurposes private data to produce insights (e.g., university researchers, the Urban Institute, data scientists).

- **Data consumer:** a person or group of people taking data insights and applying them for the public good (e.g., policymakers, practitioners).

- **Beneficiary:** some segment of the public.

Although the field is still relatively new, several distinct pathways for data philanthropy have already emerged. These pathways can be viewed as part of a spectrum of accessibility ranging from tightly restricted access to public release of data. A company considering sharing data must carefully consider its goals, legal obligations, and needs to determine the correct pathway.

Private entities, especially those governed by shareholders focused on profit, often need to prioritize their own well-being and future success, both economically and socially. For potential data providers, this often takes the form of concerns regarding data security, confidentiality, and the privacy of internal, personally identifiable information. Therefore, companies must consider their thresholds of data accessibility carefully with regard to security and privacy.

Nonetheless, the benefits of engaging in data philanthropy can be immense. Data philanthropy can help a company demonstrate good citizenship, which can improve its standing and visibility. Data philanthropy also creates a way for a company to provide greater insight on social issues that require multisector solutions. Finally, engaging in data philanthropy can help spark innovation and mitigate business risk. These benefits can provide companies with opportunities to improve the world around them while increasing their customer base and enhancing their business practices.

Further, many companies recognize that their private data holdings can complement public data and fill knowledge gaps. Private data, which may be more granular or more frequently collected than publicly available data, can support real-time decisionmaking and open new lines of inquiry. Access to these data can also save analysts time and money in the data collection stage of research. The wealth of data possessed by private companies, although not typically designed for public research, can contain

information that researchers would otherwise spend a lot of time and resources trying to acquire or replicate, if those data were available at all. By sharing their data, data providers can fulfill a vital public need at potentially little cost.

To foster stronger and more varied applications of data philanthropy, this report outlines the terrain of this emerging field. We cover several of the most common challenges data providers encounter and provide three key considerations for data providers to keep in mind when evaluating their own philanthropic potential. We also provide direct insight into data philanthropy by examining two use cases from the Urban Institute's partnership with the Mastercard Center for Inclusive Growth. We conclude with recommendations for growing a healthy and productive data philanthropy community.

The Benefits of Data Philanthropy

For Data Providers

By sharing their data, data providers can

- generate public goodwill and support community partnerships,

- provide evidence-based, data-driven insights that drive decisionmaking for social good,

- help solve public problems that cannot be addressed by existing data sources,

- align business and philanthropic activities while benefiting their customers and business ecosystem,

- mitigate potential business risks by contributing to a more informed policy environment, and

- validate their own internal data and spark innovation in their analytics processes.

In the digital age, companies are producing and compiling data more quickly than ever before. Over 2.5 quintillion bytes of data are generated every day, meaning that 90 percent of all data have been created within the preceding two years.[2] Much of this data is so-called "data exhaust," or data that is passively collected through everyday interactions with digital devices or services, including mobile phones, credit cards, and social media.[3] Data philanthropy offers an innovative, visible, and direct way for private companies to safely and securely leverage these data for the benefit of the public.

As with traditional forms of corporate giving, data philanthropy offers an opportunity for a company to accrue public goodwill. Like other more common forms of corporate giving, such as cash donations to local causes, in-kind contributions of needed supplies, or providing general operating support to charitable groups, data philanthropy allows companies to use their resources to give back to their communities and demonstrate their civic-mindedness.

However, data philanthropy can also complement more traditional forms of philanthropy.[4] In some cases, private data can yield insights not available through any other sources. In these instances, companies can provide timely resources for pressing research questions that may have no other alternative data to draw from. When a company's data are repurposed in a unique way for the public good, they can then be used to help solve social problems, improving decisionmaking and practice.

For example, by sharing a year's worth of anonymized phone data, Safaricom, one of Kenya's leading mobile service providers, helped map the spread of malaria in Kenya. Although the original data were collected for private business use, the company recognized the potential for its data to fill knowledge

gaps. Working with Harvard University researchers, Safaricom gained insight on how migration contributed to the spread of the disease (Wesolowski et al. 2012). This, in turn, translated directly into positive media attention for Safaricom on its own social media outlets and in external publications, including *Harvard Business Review* and *Medium*.[5] Safaricom provided a unique benefit to a humanitarian issue that affected the community where it does business, and the data provided by the company led directly to insights that would have been unavailable through other means.

Even in cases where the release of private data is not expected to yield immediate answers to urgent questions, providers can position themselves as trustworthy community partners. In a recent study by the Future of Privacy Forum (FPF), many companies engaged in data philanthropy voiced the opinion that by sharing data with academics, they were "demonstrating" or "unlocking" the value of their data to the public (2017).

Goodwill is not the only incentive for a company to partake in data philanthropy. Many companies struggle to align their philanthropic activities with their core business activities, and data philanthropy can act as a bridge between these two worlds. Companies can release the data they collect as a byproduct of their own business practices to philanthropic effect, and demonstrate the alignment of their profit-seeking and philanthropic goals.

As an example, Nielsen has been sharing data on individual food purchases with Feeding America since 2010 in order to combat food insecurity in the United States.[6] Using its philanthropic arm, called Nielsen Cares, Nielsen has helped Feeding America create an interactive map that visualizes key indicators of food insecurity in every state in America.[7] In using a resource collected as part of its regular business activity to address an issue facing millions of American citizens, Nielsen simultaneously directly benefited the public and increased the perception of the value of its data (Future of Privacy Forum 2017).

Data providers do not always need a firm sense of how their data could be used or what problems they could solve at the outset. Rather, they can create incentives for partners to develop innovative uses and applications of their data. Yelp stimulated innovative new uses of its data by awarding cash prizes to student papers that demonstrate "technical depth and rigor, the relevance of the results to Yelp, [their] users, or the field." This competition, now in its tenth iteration, has generated hundreds of academic papers that use Yelp's dataset.[8] Student researchers have created models that present businesses with ways to generate more profit while maintaining a positive internet presence. Yelp has not only benefited from having its name associated with public research but also spurred innovations that benefit both its own private business and other businesses participating in the Yelp ecosystem.

Finally, by sharing private data, philanthropists can spark innovation and mitigate risks for their company. Research organizations, funders, academics, and the general public are increasingly demanding both reproducibility and transparency of results, as evidenced by the growth and enhanced standards of organizations such as the Center for Open Science, the National Institutes of Health, and the Berkeley Initiative for Transparency in the Social Sciences. By making private data available to outside partners or the general public, data providers may be able to identify new trends and detect nuance collectively that is not apparent from any individual business's perspective.

Data sharing can also help validate the data. Releasing data to analysts and consumers is a type of litmus test, proving that private data used to generate business insights can withstand external academic and analytic scrutiny (Future of Privacy Forum 2017). In addition, innovation in both code and data enhancements made by the public can be integrated into the data provider's data processing pipeline to improve operational efficiency or provide new insights to management.

For example, the UN Global Pulse shared the story of a mobile carrier's private data being used to predict economic insecurity in real time. The data could provide public benefit to local policymakers, and the mobile provider also mitigates the risk of severe business cycles when early government intervention leads to more stable incomes and a more prosperous customer base.

For Data Consumers

Data sharing gives data consumers the opportunity to

- access data at potentially lower costs than many primary data collection methods,
- access more granular or more current data,
- answer new questions or existing questions faster, and
- inspire innovative solutions to entrenched problems.

Data philanthropy can help policymakers, practitioners, and researchers streamline and accelerate research for the public good. With proper due diligence and validation, data philanthropy can augment traditional forms of research data collection, fill in knowledge gaps, and provide timely answers to emerging research questions.

Collecting data for research can be costly and time-consuming. When confronted with a new research question, researchers and policymakers often face the dilemma of either collecting the data

themselves or relying on data already collected. The former enables a tailored process suited to the research questions at hand but is often costly and therefore constrained by the resources of the researcher or policymaker. The latter requires fewer resources but can mean that the data do not align perfectly with the question posed (Czajka and Beyler 2016).

Data philanthropy can provide researchers with alternatives to traditional forms of data collection. Companies are continuously aggregating valuable data that would typically be challenging for researchers to acquire or estimate on their own, and even the mundane data collected through daily business practices have the potential for broader public use.[9] On their own, datasets from a single business may paint only a partial picture, but taken with data from other businesses or combined with public data, these data can have a transformative impact.

For example, the National Institutes of Health (NIH) has organized a collective that aggregates data from the top biopharmaceutical companies to investigate better ways to diagnose and combat Alzheimer's disease.[10] Each of these companies collects a wealth of pharmaceutical data, including information on Alzheimer's medication preferences, reorder frequencies, and symptoms. The data contained by the collective houses more information than health researchers would be able to collect through more traditional means over months or even years and will hopefully shorten the time frame for potential cures.

Data philanthropists can also fill in gaps in existing knowledge by providing data that simply are not otherwise available. For example, information on daily charitable contributions is one of the major gaps in knowledge that continues to frustrate nonprofit-sector research. Researchers typically rely on data drawn from either IRS tax form data, which contain only annual levels, or voluntary disclosure of information from individual charitable organizations, which can be prohibitively costly to collect. However, the Mastercard Center for Inclusive Growth's (2016, 2017a, 2017b) *Donation Insights* reports have insights gleaned from anonymized and aggregated credit card transaction data to illuminate this very issue, providing a deeper understanding of giving trends from the donor perspective (see also the use case on charitable giving, page 23).[11]

Data philanthropy can also allow for more timely answers to research questions (Groves and Neufeld 2017; Independent Expert Advisory Group 2014). As noted above, collecting tailored research data can often be expensive and time-consuming; by the time data are released, they may be more than a year old. But private entities may be collecting relevant data on a much timelier schedule. For example, the Census releases five-year average estimates on rents and incomes by neighborhood about

one year after the final year of data is collected, but companies such as Zillow can use property-level data to track changes in prices and rents in real time (Greene and Pettit 2016).

So far, we've outlined the benefits to data providers and data consumers separately, though in practice their interests are often aligned. Companies like Nielsen, Yelp, those involved in the NIH Alzheimer's collective, and many others have already started using data to both enhance their bottom lines and benefit the public. The challenges to bringing these interests together are often technical and revolve around developing pathways for data sharing that safeguard the privacy of consumers and protect other legitimate business interests. However, as more private entities engage in data philanthropy, businesses will likely gravitate toward one of a few primary pathways that overcome these hurdles. It is to these potential pathways that this report turns in the following section.

Data Philanthropy Pathways

Data philanthropy is a new and rapidly evolving concept. A private data provider can take many avenues to unlocking its data for the public good. Based on our research and experience, current data philanthropy pathways can be characterized on a spectrum of data accessibility (figure 1).

FIGURE 1
The Spectrum of Data Philanthropy

Accessibility of Data to the Public

MORE RESTRICTED ← **1** **2** **3** **4** **5** → **MORE ACCESS**

1	2	3	4	5
Data provider mines own data for insights to share publicly	Data provider brings external researchers into own trusted network to analyze data and report findings publicly	Data provider works with other trusted data providers to form data cooperatives, aggregating different sources	Data provider shares data with trusted external researchers	Data provider allows for direct public access to datasets (often through anonymized data, data samples, and/or data tools)

URBAN INSTITUTE

For many data providers, the decision to engage in data philanthropy involves finding their place between the two ends of the spectrum: tight control versus broad access. In other words, the data provider must consider how closely it will circumscribe access to the data. Some data are more sensitive and/or might contain information that presents privacy concerns, leading the data provider to take a philanthropic pathway that restricts public access. Other data providers might recognize that their data contain less sensitive information (or take steps to anonymize their data) and choose a pathway with more open access. This report identifies five different pathways used by data philanthropists along this spectrum, beginning with the most strictly controlled and ending with the most open access. Data providers can use any of the pathways to generate useful insights; therefore, the appropriate pathways will depend on the nature of the data itself, the potential research questions to address, and the value of the analyzed results to the consumers of the data.

1. Data Provider Mines Its Own Data for Insights That Are Shared Publicly

At the most restricted end of the data philanthropy spectrum are those data providers who analyze their own data for use in public reports. The provider shares insights generated from its data but not the underlying data. In this pathway, the data provider is also the data analyst. By choosing not to release the underlying data, a company ensures that its data are kept private and secure. Data providers may still test and validate their findings by comparing them to other research or data sources or by consulting with trusted partners before making the findings public. But on this pathway, the raw data never exits the private provider's data ecosystem.

> **Examples:** In 2008, Google launched the Google Flu Trends model. Flu Trends was a data program that queried search data to track influenza outbreaks. Google analyzed this search data privately, in house, but made the findings publicly available to help health providers track flu outbreaks. The program ran until 2015. Google Flu Trends has now stopped publishing data and is donating its data to other partners in its network.[12] Google has remained a data provider but is now performing this role through a different pathway.
>
> As mentioned previously, the Mastercard Center for Inclusive Growth's (2016, 2017a, 2017b) *Donation Insights* reports have leveraged Mastercard's anonymized and aggregated transaction data to learn more about the trends in philanthropic donations, such as when donations increase or what types of organizations benefit. These insights can help philanthropic organizations become more successful by better understanding the trends in giving, such as what time of year they are most likely to receive these donations.

2. Data Provider Brings External Researchers into Its Own Trusted Network to Analyze Data and Report Findings Publicly

In this pathway, a data provider may bring certain individuals or groups of external data analysts into its network with explicit permission to work with their private data. These partnerships may be task-specific or involve lasting, consistent access to the data. Accessibility and review may still be managed by the data provider, who retains tight control over the data and its uses. However, once brought into

the network, the data analysts have access to the provider's tools and infrastructure, which are otherwise inaccessible to the public.

> **Example**: Facebook invites visiting researchers through its CommAI Visiting Researcher Program, in which researchers from universities and labs are invited to visit Facebook, collaborate on machine learning research, and produce publications and open-source code.[13] These publications and open-source applications are intended to advance the state of the art in the field and could produce broad benefits for industry, government, and research around machine learning. These partnerships are also likely to improve Facebook's internal algorithms and data analytics processes.

3. Data Provider Works with Other Trusted Data Providers to Form Data Cooperatives, Aggregating Different Sources

In this pathway, companies establish horizontal partnerships with peers in their field, often by forming a data cooperative.[14] This method presents a unique opportunity to analyze data holistically, which can help fill knowledge gaps while minimizing duplication of efforts. Pooling resources can also help address larger questions and is already being applied to such fields as justice policy and humanitarian crisis response.[15]

> **Examples**: The United Nations Global Pulse pools data from various providers to address issues related to sustainable development and humanitarian action. One of their initiatives, Data for Climate Action, currently involves nine private sector companies pooling anonymized data to propose climate change solutions to policymakers.[16] By pooling the datasets, these nine data providers will be filling knowledge gaps to create a larger and clearer picture of climate change solutions. This creates opportunities for more holistic climate change tactics.
>
> The NIH collaborative to combat Alzheimer's mentioned earlier in this report is also an example of a data cooperative, one that aggregates data from top biopharmaceutical companies to assist research devoted to Alzheimer's disease. Similar to the UN Global Pulse collective, the NIH collaborative presents opportunities to further Alzheimer's research efforts.

4. Data Provider Shares Data with Trusted External Researchers

When a company seeks to engage in data philanthropy, it may look to trusted data analysts outside its network. In these instances, rather than bring external researchers into its own secure environment, the data provider will prepare and release its data—often in the form of an aggregated and anonymized dataset—to a selected data analyst or group of data analysts. Access is still restricted, and the data provider decides which groups have access to the data and what data are shared. The analysts need not be affiliated with the data provider or have any relationship to the provider beyond the use of the data.

Examples: To better understand California's remaining water resources, Intel is partnering with researchers at the University of California, Santa Barbara, to map snow patterns in the Sierra Nevada mountains.[17] Intel is providing the researchers with over two terabytes of data per month in satellite imagery. The goal for this project, called Precision Farming, is to create a database that will enable governments and farmers to predict drought conditions and plan accordingly to increase yields and reduce crop loss.

The philanthropic partnership of Nielsen Cares and Feeding America discussed earlier in this report is also an example of this pathway.

There are many variations in how data providers manage their relationships with trusted external researchers. One method has been releasing datasets to the public via a competition format. In these competitions, the data provider puts out a call for research using its data. Gatekeeping may occur at the entry stage, where the data provider screens applicants for suitability for access, or later in the process. For example, after having demonstrated capacity with a basic dataset, the data analyst is given broader access to previously restricted-use files.

Examples: Orange Telecom, an African communications company, hosted a competition in 2013 that allowed researchers to brainstorm ideas on how to use its data to solve problems related to transportation, health, and agriculture.[18] The company rewarded the most promising idea with funding for the research project.

For many of these competitions, such as Telecom Italia's Big Data Challenge, a company may place stipulations on who can participate. On its website, Telecom Italia says that challenge participants should be "professionals, researchers, and enthusiasts." In this way, Telecom Italia restricts access before the start of the competition, ensuring that only properly qualified parties will have access to the data.[19]

5. Data Provider Allows Direct Public Access to Datasets, Often through Anonymized Data, Data Samples, or Data Tools

This most open pathway on the accessibility spectrum is defined by a data provider allowing the public free access to its data without any stipulation or barrier to entry. This method creates space for any person to interpret and manipulate the data. The data may still be anonymized and aggregated to protect privacy, and businesses that accrue large datasets may choose to share a smaller sample or summary of data.

> **Example:** As part of its AudioSet program, Google provides public access to millions of hand-labeled sound clips taken from YouTube videos.[20] The goal of the project is to improve audio event detection and work toward a full public vocabulary of sound events. Google provides a similar dataset for images under its Open Images Dataset program to improve image labeling and object detection.[21] In the future, better audio detection may be able to provide transformative services, such as helping deaf people "hear" with assisted devices, much in the same way research in computer vision and object recognition technology is beginning to help people who are blind.
>
> Twitter is engaging in this type of data philanthropy with its public Application Programming Interface (API) endpoints, in which a small sample of real-time data is released regularly to the public.[22] Twitter only requires that users create an account with Twitter and follow basic query limits. With these data, analysts can learn about how conversation topics on social media relate to offline behavior, such as civic engagement, and can track dialogue around significant events, such as natural disasters, elections, and giving days.

Each data philanthropy pathway offers its own opportunities, strengths, and challenges. The choice of which pathway to take is the result of several key considerations, which are the focus of the next section of this report.

Choosing the Appropriate Pathway

Data philanthropy can generate unique and exciting insights into a diverse array of issues of public concern, from access to quality health care or reliable and affordable transportation to adapting to climate change. However, the different starting points and perspectives of data providers and consumers create the need to establish trust and transparency throughout the process.

Data providers should think carefully about the following considerations when choosing a pathway:

- protecting privacy and ensuring data security

- minimizing transaction costs

- understanding the data and metadata

- mitigating reputational risks

Protecting Privacy and Ensuring Data Security

Data providers must ask the following questions:

- Are the data personally identifiable? Could they be reconstructed to identify someone?

- Have individuals consented to making their data publicly available?

- Are appropriate data security measures and safeguards in place?

Privacy is paramount, and data providers can work with researchers and establish their own procedures to prevent disclosure of personal information.[23] One of the first and most important steps in this process is selecting a data pathway that suits the data provider's security needs and facilitates compliance with the data provider's legal obligations. Data providers that collect personal information not explicitly intended for research purposes must choose a more restrictive pathway than providers that merely aggregate local public record data.

In these cases, data providers looking to maximize public value will likely wish to choose pathways 1 through 4—either analyzing the data themselves, bringing external data analysts within the company's firewall, creating a secure facility for data linking and access, or signing a strict memorandum of understanding and data-sharing agreement between a small number of trusted external research partners.

Data providers may also provide public, anonymized, and aggregated data, but for many research questions more granular data are often more useful. The finer the level of detail for a dataset, the wider its potential use: more granular data allow for more comparisons and/or linkages to other available data sources and may facilitate more thorough, multifaceted analysis of the dataset itself.

However, after considering the transaction costs and data understanding (discussed below), a data provider may conclude that it will provide the most public value per dollar invested by creating a public-use dataset, as Uber did with its recent release of Uber Movement data. By summarizing or aggregating information before release, the data provider can ensure that the data used by external partners do not facilitate the identification of sensitive data while still providing value to public partners.

If the dataset does not contain sensitive information, or individuals consent to making their data available, such as is the case with public Twitter accounts, the provider may consider making the data publicly available in its original form, following pathway 5. Before doing so, however, the data provider may wish to carefully consider whether the data can be linked with third-party data sources to identify individuals within the dataset.

Data providers following any pathway must ensure they have the proper process and technical safeguards in place to protect any sensitive data from unintended recipients and malicious programmers. These security processes typically include requirements that the data be stored on a password-protected computer with up-to-date software and regular password updates; that the data are stored on a confidential, encrypted disk; and that only users approved by the data provider may access the necessary files. More elaborate requirements may include two-factor authentication and separate terminals with no internet connection or USB ports.

Data providers should be wary of making system requirements needlessly complex. In some cases, these requirements can significantly increase the costs to analysts of processing the data. Providers wishing to minimize the technical burden may opt to perform any analyses in house or to provide publicly available anonymized and aggregated data.

Data providers following pathway 2 (allowing external researchers to access company data) must create a system that provides credentials to nonemployees, or onboards researchers as temporary employees, and provides access only to the systems and data necessary for research. This may require significant work if the data provider does not have similar access tiers and procedures in place for internal employees. But it is potentially less costly and time-consuming than the security requirements for following pathways 3 and 4.

Data providers following pathways 3 and 4 face the additional concerns of validating the security environments of partner organizations. In addition to the requirements listed in the previous paragraph, clearly worded and binding legal agreements, security log and audit requirements, breach notification requirements, and secure transfer sites or protocols typically satisfy most concerns. Nevertheless, data providers face increased security risks when working with external data providers and analysts (Raymond et al. 2016). In some cases, providers going this route may want to consider providing deidentified or anonymized records to their partner organizations to further minimize these risks. Data providers following these pathways should ensure they choose mature partners with a strong track record of using confidential or proprietary data and with sufficient technical staff capacity to generate the environment required.

Data providers following pathway 5 also face these same risks, though they are not as obvious at first pass. When deidentified or anonymized data are released to the public, some clever analysts may be able to identify individuals in the dataset, so companies need to be particularly careful when releasing precise count data or geographically granular estimates. Data providers choosing to follow this pathway should ensure they retain both legal and computational differential privacy experts who can advise on and quantify these risks for decisionmakers.

Minimizing Transaction Costs

Data providers must ask the following questions:

- Is there an established, trusted relationship with a data analyst?

- Is there established legal language for a data sharing or access agreement with an analyst?

- Is there an established system for reviewing the data for privacy issues before release?

- Is a technical system in place for **cleaning and validating** data, creating **aggregated and anonymized** data, **transferring data securely**, **analyzing data remotely**, **separating research data** from enterprise systems data, and **providing public access** to data?

Preparing and analyzing private data for public research requires time and financial resources. All pathways on the accessibility spectrum entail some costs to the data provider and, often, to the data analyst or consumer. However, a data provider that has some of the systems, relationships, and processes discussed below in place may find it much more cost effective to pursue one pathway over another.

Data providers pursuing pathways 1 through 3 (analyzing the data in house, providing internal access to data analysts, or sharing data via a secure facility) will face the following up-front costs: creating a system for cleaning and validating the data, separating research data from enterprise-systems data, and creating a system for reviewing privacy issues before data are released. Data providers pursuing pathway 3 will face the additional costs of establishing relationships with outside data providers or analysts and setting up a technical system for secure access and potential storage of the internal data. This additional requirement is often costly and requires a significant time investment.

Data providers seeking to work with external researchers through pathway 4 will also need to spend time and effort up front to establish a trusted relationship with external data analysts. Data providers will need to take additional time to establish legal language and data security protocols for data sharing and data access by external parties, and to set up a system for securely transferring or providing access to the data (Groves and Neufeld 2017). However, if the data provider chooses to provide raw data, it may not need to establish a system to clean and validate the data, aggregate and anonymize it, or separate research data from enterprise-systems data.

If the provider chooses to provide aggregate and anonymized data, either to data analysts or data consumers directly, it will need to establish a system to aggregate and anonymize the data to acceptable levels and additional documentation to communicate the methodology for this process. The methodology documentation is crucial to establishing a base level of trust among data consumers.

Data providers following pathway 5 (public access) will also need to consider the costs of a data transfer mechanism that works well at scale, given a much larger audience will have access. However, data providers that offer public access to the data will not need to establish a trusted relationship or legal language and data security protocols for sensitive data access.

In each pathway, the data provider will want to review the data to ensure there is no chance that private, personally identifiable information is released or identifiable in the public data. In the case of pathways 1 and 2, this may be a simple review of statistics and results produced from internal analysis, while sharing data through pathways 3 through 5 may require significantly more effort to ensure this standard is met across all groups and subgroups represented in the data.

Of course, one of the most important transaction costs to consider is communicating information about the data that may be important to analyzing the data and establishing the validity of results.

Understanding the Data and Metadata

Data providers must ask the following questions:

- Does the data provider understand potential biases in the data, including metrics like coverage rate; share of the population by different income, gender, or racial and ethnic groups; or another relevant comparison?

- Can the data provider clearly describe the variables, variable names, data collection methodology, user population, or other similar metrics?

Data collected for research purposes often come with a complete codebook describing variables, collection methods, samples, the population described, and other characteristics. In addition, there is often a substantial literature around any given dataset detailing potential biases in coverage and response rates (box 2). Private-sector data not originally collected for research may not have this level of documentation and scrutiny; if they do, this information may be scattered across the organization or poorly documented.

In pathway 1, a scattered or partially documented system may suffice, especially if a small team of internal analysts will be conducting several studies over time and can become familiar with the nuances of the data. Though these analysts will need to provide this information in a report summary and appendix to establish some base level of trust with data consumers, they will not have to produce complete codebooks and conduct detailed studies of bias and validity. The trade-off to this approach is that because only a few analysts employed by the data provider have access to the information, data consumers will likely be warier of the results produced through this pathway.

Pathways 2 through 4 involve the most effort, and the most coordination with the previous two considerations. Data providers must appropriately weigh trade-offs of increased costs both to gather this information into a single document or set of documents and to decide how much proprietary information to divulge in order to establish trust with data consumers. Organizations will have to spend time and resources on internal data experts putting their knowledge to paper in detail and test this detail with data analysts. In addition, they will need to iterate with data analysts to answer key questions of bias and validity, and these iterations will often involve both business and legal stakeholders, both to ensure a high standard of privacy and to ensure no information is released that could give competitors a potential advantage (Verhulst and Young 2017).

When sharing data with the public through pathway 5, data providers should provide detailed codebooks and information on potential dataset biases, which will involve conversations with data,

business, and legal stakeholders before release. However, data providers may not need to conduct as complete an analysis of data validity and bias as is required in pathways 2 through 4. Because the public will be able to compare the data to existing sources and the results of those efforts can be validated by independent analysts, an organization may be able to establish trust in the data's validity over time. However, because details of the individual-level anonymization and aggregation may need to remain private, these analyses may not be as valuable as those produced by data analysts granted special access to the data in pathways 2 through 4.

BOX 2
Understanding Biases through Metadata

Statistical data collected for research often follow a meticulous collection strategy that enumerates the sample size, how the sample compares to the general population, and any potential biases. Potential discrepancies between the data and the larger population are clear and explicit. However, data collected for private purposes may be biased by external factors not revealed in the data (Verhulst and Young 2017). This is particularly common when, for example, a company collects data on the people using one of its products: those data may be representative of the product users, but it may not be clear how that data may differ from others who do not use that product or use a similar competing product.

These potential biases may be obvious to the data provider, but they must also be communicated to the data analysts. The analyst may not be aware that the provider's product users tend to diverge significantly from the general population—for example, by age, race, income or gender. If the results of any analysis on these data are taken at face value, it could lead to misleading interpretations. Potential biases can include a host of different factors, including only covering a particular geography or group, being heavily weighted by another external factor, or having a limited number of reliable fields. Analysts can also help providers understand any biases in their data by comparing their data to trusted public sources, validating against their own surveys, or comparing data to studies using similar data.

Clarity of a dataset's metadata can be just as important as clarity in the data itself. This is especially true when repurposing private data for public purposes. Research datasets are often accompanied by clear, structured data dictionaries that specify and define all available fields in the data. This metadata can also include other information that relates to the construction of the dataset, such as data collection methods, sample sizes, and other pertinent information about the dataset's population.

However, private data, such as administrative data or data collected for business purposes, may not have as clearly specified metadata. If the private data are aggregated into a new dataset for public release, the metadata documentation might be scattered across various internal memos or notes (if it exists at all), rather than in a limited number of concise documents. In turn, these documents may not be appropriate for public view, as they could contain important business information.

Mitigating Reputational Risks

Data providers must ask the following questions:

- When producing public data, does the company have a method to educate users, language that must accompany any use or analysis, or restrictions on data use?

- When partnering with others, have the various parties disclosed their reputational risks to each other, with specific examples? Common questions to consider when assessing risk include the following:

  » Does the data provider allow the data analyst to publish without editorial review from the data provider?

  » Do the data analyst and consumer understand when to clearly convey when data have been collected by the provider and when they are additions or estimations from additional data and modelling?

  » Do the data provider and analyst understand which research topics may be potentially harmful to each party and require additional conversation?

  » Has the provider properly explained which pieces of information may be disclosed before the research process begins?

Data philanthropy creates reputational, brand, and business risks for the data provider and analyst. These risks must be properly mitigated to unlock the potential of private data for the public good.

For all pathways on the accessibility spectrum, data providers will face reputational risks to their organization. The three primary risks they face are the perception that the provider is collecting or sharing sensitive data that it is not, the perception that the research topic or data convey a particular political or religious viewpoint, and the disclosure of information that is either proprietary or potentially valuable to industry competitors.

Data providers choosing to analyze the data themselves will have to manage these risks internally, while data providers pursuing pathways 2 through 5 face additional risks that they should consider carefully before entering into these arrangements. The provider should have a process, restricted data portal, terms of use, or license in place to educate the data analyst as to the language necessary when disclosing which data were collected by the provider, which parts of the provider's data collection and preparation process can be discussed, and any restrictions on data use.

When working with external researchers or partners in pathways 2 through 4, data providers should consider that these data analysts also face reputational risk; analysts wish to be perceived as

independent scientists, and they may require that they be allowed to publish findings without editorial review from the data provider. Providers will want to ensure that they allow for privacy and legal review for publications and that they allocate time earlier in the process to carefully select research partners and topics.

Companies that require editorial control over research, or who face large reputational risks, may be safer choosing pathway 1—or, if possible, pathway 5—with an appropriate restricted data portal, legal terms of use, and licensing requirements. However, companies that are comfortable mitigating the reputational risks associated with pathways 2 through 4 may find that the data provided create more impactful research and results because of the independent nature and externally validated scientific rigor research partners can offer.

Each of these considerations is vital for private companies when judging the appropriate data philanthropy pathway for their interests. Although these considerations must be taken seriously, different pathways are available for all interested data providers to engage in data philanthropy, regardless of the outcome of these considerations.

Use Cases

To illustrate private data's potential as a source of important information for policymakers, practitioners, and researchers, this section showcases one example of data philanthropy in action: the partnership between the Urban Institute and the Mastercard Center for Inclusive Growth ("the Center"). Through this partnership, Urban explored how Mastercard's insights could be used to understand several key policy issues related to charitable giving and inclusive redevelopment in US cities. To shed light on these important matters, the Center provided Urban access to a set of Mastercard's transaction data narrowly limited in both time and geography, which was both aggregated and anonymized, as well as subject to numerous additional privacy and data protection safeguards. We refer to the dataset as "Mastercard data insights." We present the findings from these explorations in the two use cases that follow.

Although the results are preliminary, these use cases highlight how data philanthropy can fill knowledge gaps; the use cases also illustrate some of the considerations raised in this report. The partnership between the Center and Urban is an example of pathway 4, **a data provider sharing data with a trusted external researcher.** As noted earlier, each pathway and partnership comes with its own considerations.

Ensuring compliance with privacy and data protection laws, standards, and other requirements— including those relating to information security—is significant to both Urban (the data analyst) and the Center (the data provider). To this end, the Center and Urban entered into a comprehensive data license agreement that subjected the partnership generally and Mastercard data insights specifically to robust controls. Throughout this partnership, the Center and Urban worked together to minimize transaction costs through regular meetings, communication, and information sharing.

As one example of a privacy and data protection control, the Center indexed the Mastercard data insights through an internal system reflecting an undisclosed—but constant—base. In other words, all information presented below, unless otherwise stated, does not reflect *actual* dollars received or *actual* number of transactions, but a *constant index* calculated by the Center. This practice, while preventing any direct comparison to external financial measures, still allows for proportional and relative comparisons across Mastercard data insights.

Because of the exploratory nature of our work, the analysis did not have an extensive literature of existing public resources or documentation to refer to, and the potential public uses of these insights were largely uncharted. Therefore, a particularly important consideration during this partnership was

understanding the data and metadata (see page 18). Urban had to establish a baseline understanding of any characteristics of the underlying data relating to these insights that might influence the interpretation of the results. For example, Urban had to understand whether there may be any differences among and between Mastercard and non-Mastercard users that would have a significant impact on the results of the analysis.

At the end of 2016, Mastercard held 31.6 percent of the market share of credit cards by network purchase volume.[24] Therefore, the information captured by Mastercard represents a large proportion—but not the entire population—of credit card holders. Further, research by the Federal Reserve Board suggests that the demographics of all credit card holders in the United States skew—for example, with respect to affluence—in a manner that diverges compared with the country's population as a whole (Larrimore et al. 2017). These factors must be taken into consideration in any research conducted using credit card transaction data, meaning that results may not reflect spending patterns of the general population or of those using other forms of payment, such as cash.

After reviewing the Mastercard data insights, researchers at Urban were able to glean insights on charitable giving patterns and on the equitable development activities around the future 11th Street Bridge Park project in Washington, DC.

Charitable Giving Use Case

Detailed information on individual charitable giving remains one of the most significant knowledge gaps for the nonprofit sector. The most common sources of information on charitable contributions are based on publicly available tax data on contributions reported by nonprofits, which can be found on the Internal Revenue Service's Form 990 (the annual financial assessment required by all tax-exempt organizations filing in the United States). However, data sources like Form 990 lack granularity and timeliness: they cannot speak to subjects like the number of transactions going to a specific cause or about short-term spikes or downturns in charitable giving.

The Mastercard data insights could fill these critical knowledge gaps and improve on existing data sources. First, existing comprehensive tax data often take a year or more to be processed, prepared, and made available for public consumption. Access to financial transaction data could reduce or eliminate this lag, allowing for much greater responsiveness to events impacting the charitable sector.[25] Second, the data provided by the IRS only includes information on charitable organizations required to publicly report their finances. The Mastercard data insights include information on organizations not required to

file with the IRS, such as small nonprofits. Third, the Mastercard data insights include information not reflected in other sources, such as number of charitable transactions and aggregated daily giving.

In this use case, we test whether Mastercard's data insights can help fill these knowledge gaps and yield new information. To do this, we first benchmark the Mastercard data insights against two other sources of nonprofit data. Then, we examine the novel benefits of the Mastercard data insights through two unique opportunities: a look at average donations in the nonprofit sector and an analysis of daily giving patterns.

How Representative Are the Mastercard Data Insights?

Given that the Mastercard data insights represent a new source of information for the nonprofit sector, the first step for our analysis was to better understand any potential biases in the data through a comparison to other available sources of nonprofit sector data. We assessed Mastercard data insights against two other well-known and respected data sources: Giving USA's estimates of charitable giving, and the National Center for Charitable Statistics's (NCCS) aggregations of charitable contributions reported by public charities.

We began by calculating the proportion of indexed total contributions[26] received by nonprofit organization type in 2015,[27] the most recent year available for each source of data. The results are presented in figure 2. For most subsectors, the share of total giving was similar between the different data sources. The largest discrepancies between Mastercard data insights and the other two sources were in the health and human services subsectors.

FIGURE 2
Share of Total Giving by Subsector, 2015
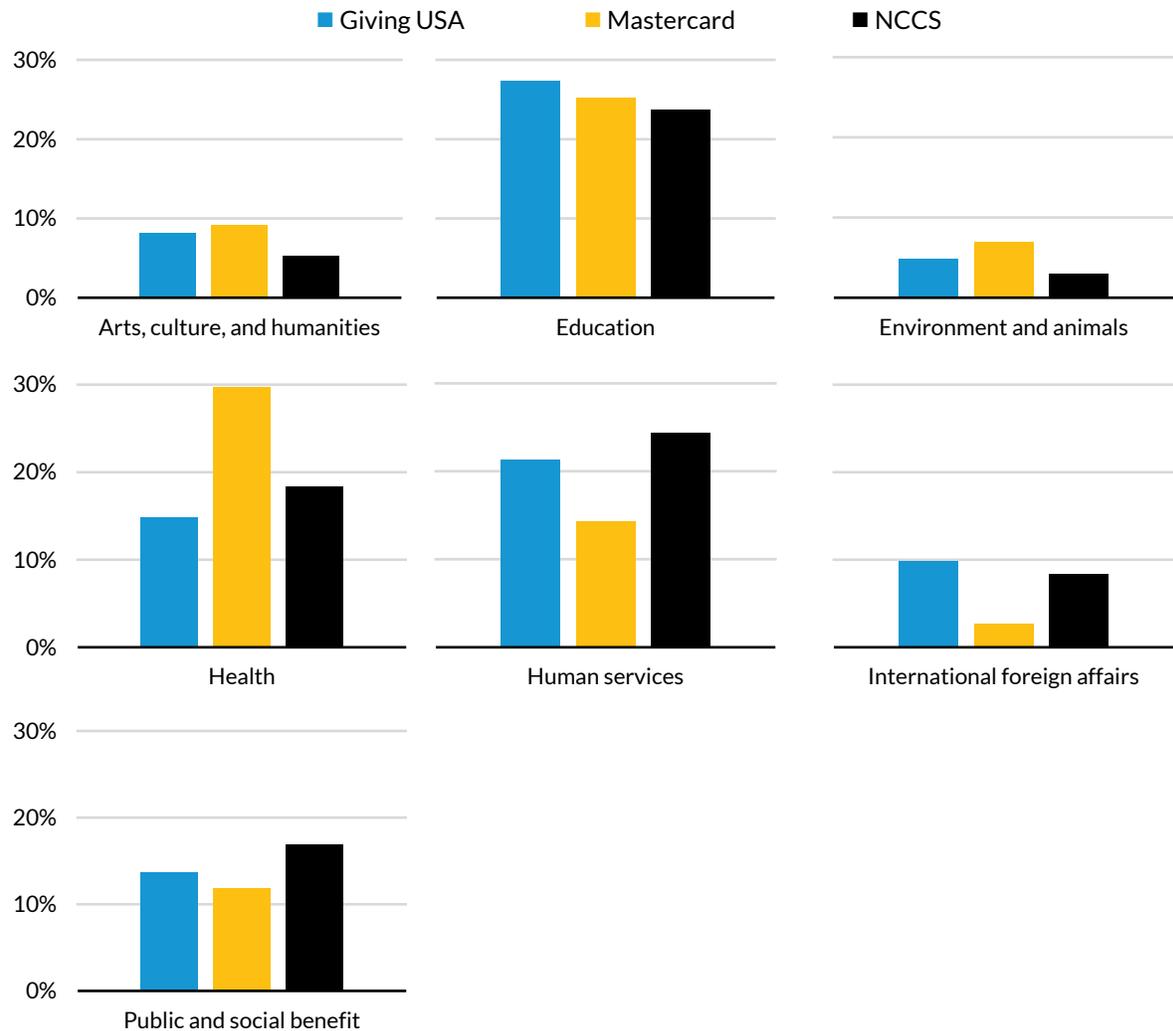
Legend: ■ Giving USA ■ Mastercard ■ NCCS

Subsectors shown: Arts, culture, and humanities; Education; Environment and animals; Health; Human services; International foreign affairs; Public and social benefit.

Sources: Urban Institute, National Center for Charitable Statistics, Core Files (Public Charities 2015); Mastercard data insights; Giving USA Foundation, *Giving USA: The Annual Report on Philanthropy for the Year 2016* (Chicago: Giving USA Foundation, 2017). **Notes**: NCCS = National Center for Charitable Statistics. Subsector definitions vary from the Mastercard Center for Inclusive Growth's Donation Insights report.

These discrepancies could be caused by slight differences in what the underlying data capture. Giving USA and NCCS both use publicly available tax information as the starting point for analysis. This tax information aggregates nonprofit revenue designated as charitable contributions from all sources (individuals, corporations, foundations, and others) through all forms of tender (cash, check, credit cards, etc.). The Mastercard data insights, on the other hand, are drawn from all electronic payments to

nonprofit recipients. As such, they exclude other (nonindividual) sources of giving or nonelectronic forms of tender, and they include transactions that may not be classified as contributions by the recipient organization (e.g., fees for services or other types of revenue). Consequently, subsectors that rely more heavily on electronic payments, giving from individuals, or fees for services may be represented more heavily in the Mastercard data insights than in Giving USA or NCCS data.[28]

We also explored the possibility that the Mastercard data insights may differ from other sources of data when comparing change over time. We calculated the proportional growth of each subsector from 2011 to 2016 using the subsector's 2011 total contributions as the starting point for each data source.[29] Once again, despite small differences, most subsectors displayed similar growth across data sources, while the health subsector showed greater divergence between the Mastercard data insights and our two benchmark data sources (figure 3). Additionally, the Mastercard data insights reflected much higher growth for the human services subsector than either NCCS or Giving USA. This difference suggests that the growth in card-based giving to human services may be outpacing the growth from other sources such as cash, grants, or bequests.

FIGURE 3

Percentage Growth by Subsector, 2011–16
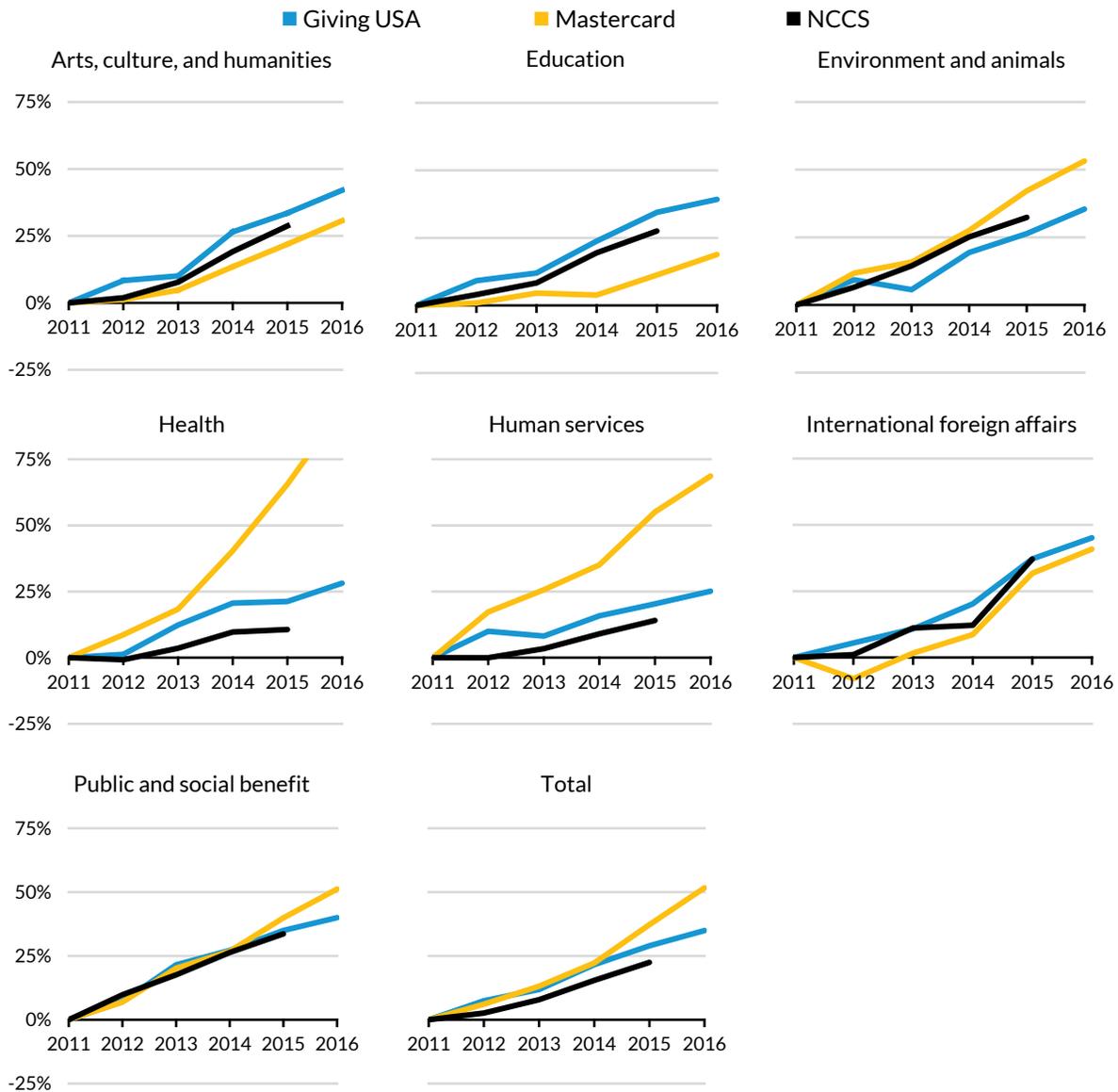
Legend: ■ Giving USA   ■ Mastercard   ■ NCCS

Charts: Arts, culture, and humanities; Education; Environment and animals; Health; Human services; International foreign affairs; Public and social benefit; Total

Sources: Urban Institute, National Center for Charitable Statistics, Core Files (Public Charities 2011–15); Mastercard data insights; Giving USA Foundation, *Giving USA: The Annual Report on Philanthropy for the Year 2016* (Chicago: Giving USA Foundation, 2017).
Notes: NCCS = National Center for Charitable Statistics. The most recent year available for NCCS data is 2015. Subsector definitions vary from the Mastercard Center for Inclusive Growth's Donation Insights report.

In conclusion, the Mastercard data insights do show some variation when compared against other data sources of giving. However, these totals are within plausible bounds, given differences in collection and aggregation of data. The understanding of these potential biases will help inform further analysis, and the divergences may be grounds for future analysis on spending patterns among credit card holders.

Opportunity 1: Measuring Average Contribution Level

In addition to allowing cross-validation of assumptions made from other sources of charitable giving data, the Mastercard data insights also provide unique information not available through those other sources, such as the average contribution amount per transaction. Existing sources of tax data provide only aggregated amounts of contributions by individuals but offer little insight into the number of contributions. For example, *The Nonprofit Almanac* combines publicly available tax data with information on personal giving to estimate the average charitable contributions per tax filer (approximately $1,660 in 2013) (McKeever, Dietz, and Fyffe 2016). However, this tells us little about the number of causes a person might donate to or the size of a single discrete charitable contribution.

The Mastercard data insights, on the other hand, provide information on the total (indexed) amount of contributions as well as the *number* of contributions. At the most basic level, this allows for a calculation of average contributions by nonprofit type, as seen in figure 4 below.

FIGURE 4
Average Contribution Size by Subsector, 2016
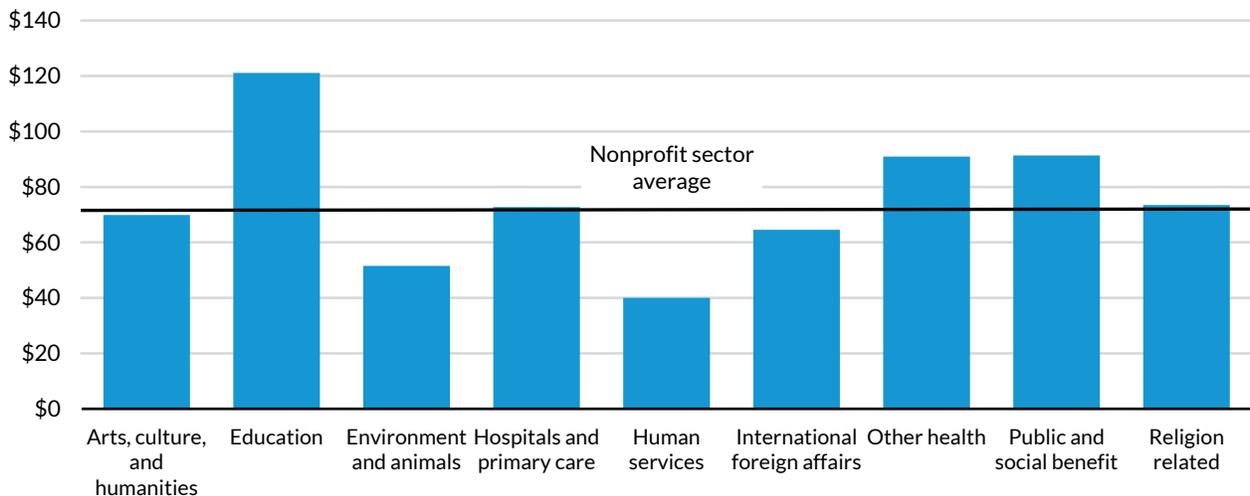
Source: Mastercard data insights.

The Mastercard data insights reveal that the education subsector received the highest average donation amount per contribution and the human services subsector received the lowest.[30] Again, the Mastercard data insights do not distinguish between payment method or type, so the education subsector may be weighted by high-cost transactions made electronically, by individuals, or for services other than contributions.

In our analysis, the human services and environment and animals subsectors consistently had the lowest and second-lowest averages, respectively, for every year of available data (2011–16). The low averages may point toward differences in donation vehicles for these sectors.

Recurring payments may also be contributing to the low averages for these two subsectors. The Center's *Donation Insights* report notes that recurring payments have been growing as a share of the nonprofit sector since 2009, but average recurring payment sizes for human services and environment and animals organizations are both below average (Mastercard Center for Inclusive Growth 2017a). By following these trends in charitable giving patterns, the Mastercard data insights can provide information beyond the simple aggregations available through typical data sources.

Opportunity 2: Measuring Daily Giving Patterns

Because of the frequency of data collection, Mastercard's data insights can speak to charitable giving during short-term events in a way that sources that rely on annualized giving data cannot.

For example, the Mastercard data insights can illuminate giving on so-called "giving days." A new form of giving led in large part by local and regional community foundations, giving days focus on targeted fundraising within a limited time span. The largest giving day event in the US is Giving Tuesday, which began in 2012 as a philanthropic response to the commercialization of the American holiday season and raised an estimated $274 million in 2017.[31] The rapid growth of Giving Tuesday has been lauded as a successful fusion of technology and philanthropy, with well over 7,000 nonprofits participating in recent years.[32]

Mastercard data insights allow us to track the growth in donations on Giving Tuesday in greater detail by disaggregating the information by subsector. As seen in figure 5 below, in 2012, the amount given to the education subsector eclipsed all other subsectors by far. However, giving toward human services organizations on Giving Tuesday has increased each year, culminating in 2015, when it surpassed the education subsector to receive the highest total value of transactions that year. This

increase may represent the growing attention received by Giving Tuesday and the event's strong partnership with human services organizations.

Giving Tuesday Donation Amounts by Subsector, 2012–15
Indexed amount of transactions
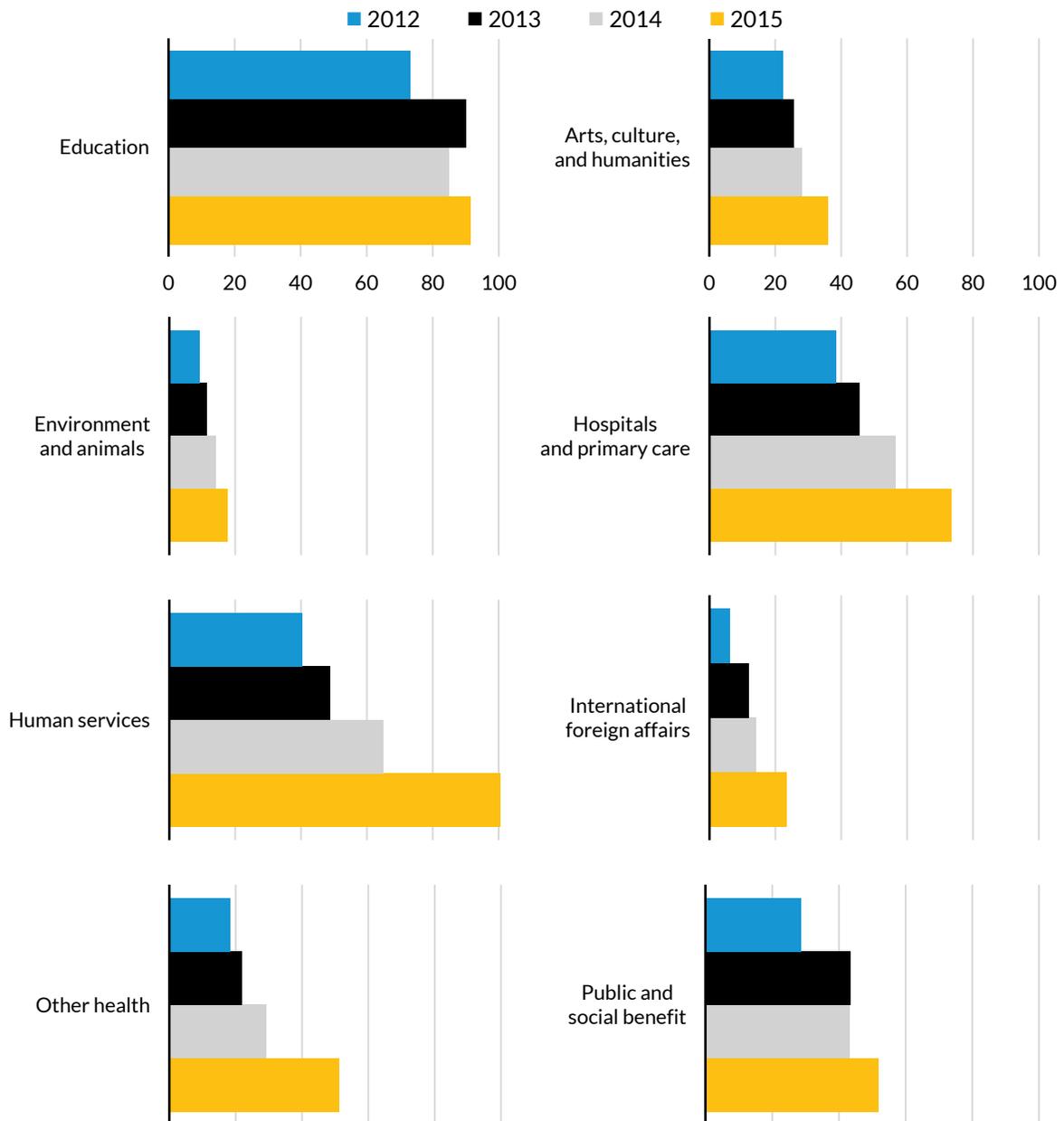
Source: Mastercard data insights.
Notes: Data for 2016 excluded because of missing data. **Religion related** category excluded because of missing data.

Mastercard data insights could provide similar understanding about the dozens of similar giving days that have emerged in recent years, representing a broad range of causes and communities. The data also have the potential to clarify the effects of other short-term events on the charitable sector. For example, the Center's *Donation Insights* reports have provided greater understanding of the impact of natural disasters on charitable giving.[33] The potential for researchers to explore giving responses to particular short-term events would greatly expand the sector's knowledge.

Conclusion

With so much interest and debate regarding charitable giving in the wake of the recent Tax Cuts and Jobs Act, which reduces the incentive to use the charitable deduction for many families, timely and granular data on charitable giving are more important than ever before. Mastercard's insights have the potential to help us better understand shifts in patterns of giving, which in turn can better inform policy and practice.

This use case revealed that although the Mastercard data insights differ slightly from other widely used sources of information on charitable giving, the basic trends largely conform to known patterns of giving. As shown here, our analysis of Mastercard data insights produced results similar to analyses of sources derived from publicly available tax data. Compared with tax data, the higher frequency of Mastercard's data collection could allow future researchers to reduce the delay in tracking trends in charitable giving. Although it is essential to recognize the differences between Mastercard data insights and other sources of data to ground future analysis, the similarities shown here illustrate the potential for Mastercard data insights to complement existing public sources of information on charitable giving.

In this piece, we also touched on how Mastercard's data insights can replicate and expand nonprofit sector knowledge on individual charitable giving by providing information not currently available. We saw how Mastercard information on number of charitable contributions, in addition to total amounts, can be used to assess charitable giving by subsector and highlight trends in charitable giving vehicles. We also explored one example of how Mastercard data insights can speak to giving related to short-term events with a degree of specificity publicly available tax information cannot.

Thus, the potential for private data to fill wide gaps in our understanding of charitable giving is immense. Data already being privately collected can help answer these and other nonprofit sector questions. As more organizations engage in data philanthropy, we will only continue to expand our knowledge and comprehension of the nonprofit sector.

Neighborhood Equitable Development Use Case

Local redevelopment projects designed to repurpose underused urban land or structures, such as New York City's High Line and the Atlanta Greenbelt, have the potential to connect communities and allow all residents to benefit from improvements. However, without broad participation in agenda setting, intentional strategies and goals, and accountability mechanisms, these projects can contribute to runaway gentrification or exacerbate racial and economic inequalities in their neighborhoods. Unfortunately, data to help understand the impact of these projects in real time and at the neighborhood level are rare.

The 11th Street Bridge Park in Washington, DC (Bridge Park) is an example of a local redevelopment project that prioritized equity and inclusion at the outset. The Bridge Park concept was launched in 2011 by the Office of Planning of the District of Columbia. Slated to open by late 2019, the project will repurpose an obsolete highway bridge spanning the Anacostia River into an elevated space for recreation, environmental education, and the arts. When completed, the Bridge Park will connect Capitol Hill—one of the city's more privileged communities—with neighborhoods east of the Anacostia River that have suffered from chronic underinvestment.

Although the park (and other projects planned or underway) could bring much-needed investment to neglected neighborhoods east of the Anacostia River, new development also raises the specter of gentrification and the displacement of existing residents because of rising costs of living. These risks are very real to many people, particularly people of color, who have experienced past waves of development and displacement.[34]

To respond proactively to these concerns, Bridge Park has engaged with communities on both sides of the river and partnered with the Washington, DC, office of the Local Initiatives Support Corporation to plan for equitable development around the new project. In 2015, Bridge Park convened an equitable development task force of local officials, researchers, planners, and community experts. The task force's recommendations were compiled into a formal plan consisting of three main pillars: housing, workforce development, and small business. The plan outlines steps the Bridge Park can take, either alone or in partnership with other organizations, to promote inclusive development in neighborhoods adjacent to the new park that benefits people and businesses currently in those areas.

Role of Data

The Bridge Park project and its partners will use data when implementing the equitable development plan to help frame the initial planning and execution and to ensure accountability, both internally and externally. The Bridge Park partnered with the Urban Institute to create a data framework for equitable development and to begin compiling data to measure baseline conditions and track progress toward achieving the plan's goals.[35] Urban has an extensive collection of local data related to housing and, to a lesser extent, workforce and labor issues, but data on small businesses are much harder to come by. Business data are primarily compiled by private vendors, who charge fees and put restrictions on their use. Further, these sources often lack data on business volumes or customers that can be used to assess the quantity and quality of business activity in a community.

BOX 3

11th Street Bridge Park Small Business Development Strategies

- Support and nurture a thriving network of small businesses that operate on the Bridge Park.

- Leverage the Bridge Park to build and sustain small businesses in the surrounding community.

- Ensure the Bridge Park is deeply connected to business corridors on both sides of the Anacostia River.

Source: 11th Street Bridge Park equitable development plan.

We looked at whether the Mastercard data insights could answer questions regarding business activity in neighborhoods that might help inform equitable development activities. Urban explored two primary types of questions using these data:

- **To what extent does retail leakage occur in the project's impact neighborhoods and how does that vary by type of business?**

  This question addresses how much consumers living in the areas near the park spend at local businesses as opposed to other businesses inside or outside Washington, DC (retail leakage). Significant retail leakage might suggest opportunities to invest in existing local businesses or create new businesses to meet consumer needs.

  By estimating residential zip codes (as discussed below), Mastercard data insights also have the potential to shed light, at least indirectly, on questions about displacement.

- **To what extent do cardholders change their residential location from year to year?**

  This question provides an estimate of residential displacement, which may be voluntary or involuntary, and can be used to monitor whether displacement rates differ in the two impact neighborhoods and are constant or increase or decrease over time. The answer to this question could help shed light on the larger antidisplacement goals of the equitable development plan.

Understanding Retail Leakage

Understanding where consumers spend their money is an important data point for local business development strategy. For this analysis, Mastercard provided access to anonymized and aggregated data of total relative spending at businesses based on a residential zip code estimate for cardholders. Business locations were summarized based on census tracts situated east of the Anacostia River (the lower-income community) and west of the river (the higher-income community).[36] Estimated residential cardholder locations were based on the respective zip codes roughly corresponding to the same east and west areas defined for businesses.

TABLE 1

Share of Spending on Businesses on East and West Sides of the River (All Industries), 2012–16

By cardholders from estimated zip codes east and west of the river

Estimated cardholder residential zip code	Businesses east of the river	Businesses west of the river
Combined areas	27	73
East of the river	63	37
West of the river	20	80

Source: Mastercard data insights.

Table 1 examines spending by cardholders estimated to live in zip codes east and west of the river at businesses in each of those respective areas. The analysis suggests consumers generally shop at businesses closer to where they live. For cardholders east of the river, about two-thirds of their local spending (63 percent) is at businesses east of the river. For cardholders west of the river, 80 percent of their local spending is at businesses west of the river.

Although these numbers varied somewhat across different industry groups, such as restaurants or beer, wine, and liquor establishments, the overall pattern of local spending was fairly consistent for cardholders west of the river. Cardholders east of the river seemed more likely to spend at businesses west of the river in certain categories, however, such as retail apparel, computers and electronics, home

improvement, and sporting goods. This pattern likely reflects the relatively sparse supply of retail businesses on the east side of the river.

Understanding Residential Displacement

The second potential use of the Mastercard data insights is in tracking relative rates and trends of residential mobility within particular zip codes. Intense economic development may result in voluntary or involuntary residential displacement, which has been very hard to measure using existing data sources. Mastercard leveraged its estimated residential cardholder zip codes and provided access to aggregate summary statistics for cardholders living east or west of the river in 2014 (figure 6). For this analysis, we looked at zip codes east and west of the river to create a baseline for future analysis using the insights as a proxy for what percentage of cardholders likely remained in the city or moved elsewhere (based on changes in the residential zip code estimate) over the course of the year.

FIGURE 6
Annual Residential Mobility Rates (Percentages) for Cardholders East and West of the River, 2014

■ Cardholders east of the river (zip 20020)   ■ Cardholders west of the river (zip 20003)
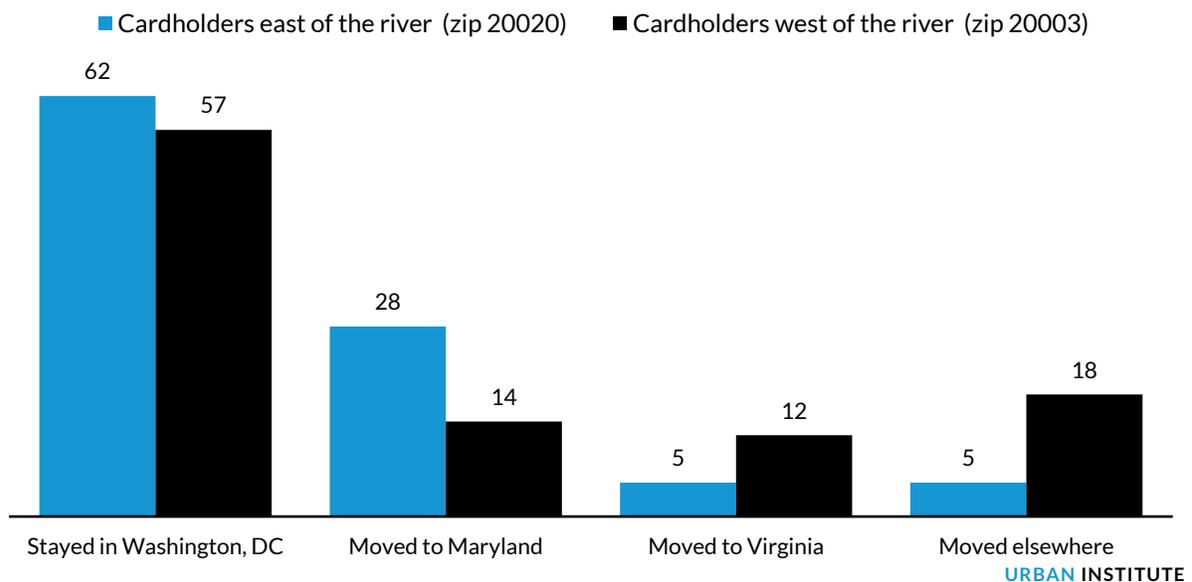
Source: Mastercard data insights.

Overall, the rates at which cardholders in the two zip codes remained in Washington, DC, were not very different. Sixty-two percent of cardholders east of the river and 57 percent of cardholders west of the river likely remained in Washington, DC. But the likely destination locations of the movers differed. Cardholders east of the river were twice as likely to move to Maryland; conversely, cardholders west of

the river were more likely to move to Virginia or elsewhere in the US. Mastercard also provided access to these same insights for 2012–13, which showed nearly identical patterns.

Conclusion

The Mastercard data insights have the potential to help us understand the changes going on in neighborhoods where development is taking place and can inform activities to address those changes. The data can reflect some interesting patterns in retail spending and estimated residential movement that other data cannot. In addition, these data are collected frequently, which can be helpful in gathering immediate insights in rapidly changing neighborhoods.

Making analyses informed in part by Mastercard data insights accessible to local government agencies, community development organizations, and residents could provide valuable intelligence on local changes affecting the retail landscape. The use of these data should be explored further to determine where they can provide the most value.

Paths Forward

As seen in the use cases above, private data can contribute to significant and meaningful progress toward answering public questions. Urban's own partnership with the Mastercard Center for Inclusive Growth demonstrated how a data provider can provide information not available through other means. In the cases presented above, we showed how Mastercard's data insights could be applied to studies of charitable giving and inclusive urban redevelopment projects: two disparate social policy subjects where salient research questions are often preemptively stymied by the lack of granularity, timeliness, or specificity of publicly available data. In both instances, Mastercard data insights were able to address critical knowledge gaps, furthering our understanding.

This partnership illustrated how private data held by a single data provider can have broad application to public research conducted by a trusted external research partner. But this relationship represents only one of the potential pathways of data philanthropy. Data providers mining their own data for insights, forming horizontal data cooperatives, and releasing anonymized and aggregated versions of their data to the public can all contribute to the public good.

The data philanthropy field is broad, diverse, and growing. As more data providers enter the field, the applications of private data will grow. And as data philanthropy continues to evolve, there are several things that data providers can do to leverage their data assets for the public good.

First, data providers should recognize that they operate within a larger community of data philanthropists. As shown above, the independent contributions of any single data provider can fill vital knowledge gaps across broad research fields: we have seen how administrative telecom data has helped track the spread of deadly diseases, how satellite imagery can help farmers predict and plan for droughts, and so on.

Yet in addition to these tremendous individual offerings, organizations are increasingly taking advantage of their shared experience with data philanthropy to tackle even larger concerns. Collaboratives such as UN Global Pulse are creating bridges between different data providers working on the same issue. Bloomberg's Data for Good Exchange has brought industry data scientists together with academics, governments, and nongovernmental organizations to work on social problems.[37] And Mastercard's Center for Inclusive Growth has called for other data providers to share their data, knowledge, and expertise.[38]

Bringing together independent organizations to address the same problems benefits both data consumers and providers. Potential consumers reap the benefits of aggregated datasets that surpass those that any single provider could contribute. Data providers benefit from these partnerships as well. A better understanding of efforts across the field can help providers identify key knowledge gaps they may be able to fill while minimizing potential costs associated with duplicating efforts. In addition, providers' business analytics processes may benefit, and organizations may expand their potential customer bases through the impact of their efforts.

Therefore, cultivating a data philanthropy community of practice benefits both data providers and data consumers, and data philanthropists stand to gain more by working cooperatively than they could achieve independently.

Second, most data providers decide to engage in data philanthropy after considering several key factors. Potential data philanthropists must take into account elements unique to their own organization's structure, and these elements may lead to very different philanthropic pathways. One organization's privacy and data protection obligations, support capacity, and transaction costs might encourage it to take a different path than another organization operating in the same field but without the same constraints.

Although the resulting pathway may vary across organizations, the key considerations are similar for many data providers, and potential data philanthropists need not reinvent the wheel. By creating shared principles related to best practices for access, privacy, and data validation, data providers can minimize planning costs for the data philanthropy community. By establishing a standard foundation for data philanthropy, providers can encourage entry into the space. More experienced providers should consider creating open-source templates or platforms to encourage the spread of data philanthropy.

Third, the range of data philanthropy pathways means that data providers have many different potential models to choose from. Although an organization may consider its internal data holdings to have wide public merit, moving directly to broad open access can be daunting. The release of private data to the public requires careful planning and deliberation and appropriate safeguards.

Data providers interested in engaging in data philanthropy should consider using trusted data intermediaries to assess their broader potential before moving directly to public open access. This report has noted several possible forms for this: bringing external researchers into the organization's trusted network, partnering with external experts, or hosting competitions to assess potential partners. By restricting access to trusted partners, data providers can assess the broader appeal of their holdings while mitigating risk. This can be especially productive if the data provider finds external experts

knowledgeable about the particular public research or policy field the provider is looking to engage with. By bringing an outsider's perspective, the data intermediary may be better positioned to assess knowledge gaps, bias concerns, metadata, or other elements. If the process is productive for both parties, the data provider can consider moving future releases to a more open pathway on the accessibility spectrum.

Finally, the fast development of data philanthropy means there is still plenty of room for innovation. Potential philanthropists should not feel constrained by the approaches currently available, which should be treated as launching pads for further exploration. For example, organizations limited by privacy and data protection obligations could consider developing synthetic public datasets that protect, among other things, privacy and confidentiality while still showcasing the breadth of the available data. Or organizations that prohibit access to granular data might create a platform for researchers to submit code against the confidential dataset to get results.

This report has been intended as a guide to a nascent but exciting and promising field. We describe the current state of data philanthropy and suggest where it might be going and offer key considerations for data providers to keep in mind. Nonetheless, data philanthropy is still a largely uncharted field, and even organizations taking the same pathway may have different experiences. There is still considerable space for new participants and new ideas. It is an exciting time for data philanthropy, an emerging practice that has much to offer data providers and consumers alike. Together, all participants can unlock the power of private data for public good.

Notes

1. Randy Bean, "Mastercard's Big Data For Good Initiative: Data Philanthropy On The Front Lines," *CIO Network* (blog), *Forbes*, August 7, 2017, https://www.forbes.com/sites/ciocentral/2017/08/07/mastercards-big-data-for-good-initiative-data-philanthropy-on-the-front-lines/; Andreas Pawelke and Anoush Rima Tatevossian, "Data Philanthropy: Where Are We Now?" *United Nations Global Pulse* (blog), May 8, 2013, https://www.unglobalpulse.org/data-philanthropy-where-are-we-now; Stewart Rassier, "Data Philanthropy: The High-Tech Future of Social Impact," *Corporate Citizenship Perspectives* (blog), Boston College Center for Corporate Citizenship, August 30, 2016, http://corporatecitizenship.bc.edu/data-philanthropy; Dave Tang, "Data Philanthropy: Bridging Tech and Social Good," *Content Tracks* (blog), Datacenter Dynamics, April 3, 2017, http://www.datacenterdynamics.com/content-tracks/servers-storage/data-philanthropy-bridging-tech-and-social-good/98095.fullarticle.

2. "Data Never Sleeps 5.0," Domo, accessed May 16, 2018, https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1.

3. "Big Data for Sustainable Development," United Nations, accessed May 16, 2018, http://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html.

4. Matt Stempeck, "Sharing Data Is a Form of Corporate Philanthropy," *Harvard Business Review*, July 24, 2014, https://hbr.org/2014/07/sharing-data-is-a-form-of-corporate-philanthropy.

5. Stefaan G. Verhulst, "Mapping the Next Frontier of Open Data: Corporate Data Sharing," *Medium* (blog), January 5, 2015, https://medium.com/internet-monitor-2014-data-and-privacy/mapping-the-next-frontier-of-open-data-corporate-data-sharing-73b2143878d2; Harvard T.H. Chan School of Public Health, "Using Cell Phone Data to Curb the Spread of Malaria," October 11, 2012, https://www.hsph.harvard.edu/news/press-releases/cell-phone-data-malaria/.

6. Nielson, "Nielsen and Feeding America Collaborate to 'Map the Meal Gap'," September 20, 2015, http://sites.nielson.com/newscenter/nielsen-and-feeding-america-collaborate-to-map-the-meal-gap/.

7. "Food Insecurity in The United States," Feeding America, accessed May 18, 2018, http://map.feedingamerica.org/.

8. "Yelp Dataset Challenge," Yelp, accessed May 16, 2018, https://www.yelp.com/dataset/challenge.

9. Gideon Mann, "Private Data and the Public Good," *Medium* (blog), May 17, 2016, https://medium.com/@gideonmann/private-data-and-the-public-good-9c94c656ff28.

10. "Accelerating Medicines Partnership (AMP)," National Institutes of Health, accessed May 16, 2018, https://www.nih.gov/research-training/accelerating-medicines-partnership-amp.

11. "New Insights on Charitable Giving," Mastercard Center for Inclusive Growth, December 14, 2016, https://mastercardcenter.org/insights/new-insights-charitable-giving/.

12. The Flu Trends Team, "The Next Chapter for Flu Trends," *Google AI Blog*, August 20, 2015, https://research.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html. For more on Google Flu and the efforts made to refine the model, see Butler (2013), Lazer and colleagues (2014), and Donna Marbury, "Google Flu Trends collaborates with CDC for more accurate predictions," *Medical Economics*, November 5, 2014, http://www.medicaleconomics.com/health-care-information-technology/google-flu-trends-collaborates-cdc-more-accurate-predictions.

13. "Visiting Researchers and Postdocs," Facebook Research, accessed May 16, 2018, https://research.fb.com/programs/post-docs-and-sabbaticals/.

14 Stefaan G. Verhulst and Andrew Young, "How the Data That Internet Companies Collect Can Be Used for the Public Good," *Harvard Business Review*, January 23, 2018, https://hbr.org/2018/01/how-the-data-that-internet-companies-collect-can-be-used-for-the-public-good.

15 The White House Office of the Press Secretary, "Fact Sheet: Launching the Data-Driven Justice Initiative: Disrupting the Cycle of Incarceration," June 30, 2016, https://obamawhitehouse.archives.gov/the-press-office/2016/06/30/fact-sheet-launching-data-driven-justice-initiative-disrupting-cycle; GSMA, "GSMA Launches Big Data for Social Good Initiative," February 27, 2017, https://www.gsma.com/newsroom/press-release/gsma-launches-big-data-for-social-good/.

16 "Data for Climate Action," United Nations Global Pulse, accessed May 16, 2018, https://www.unglobalpulse.org/data-for-climate-action.

17 "Mapping Snow Melting in the Sierra Nevada," Data Collaboratives, accessed May 16, 2018, http://datacollaboratives.org/cases/mapping-snow-melting-in-the-sierra-nevada.html.

18 "The D4D Challenge is a great success!" Orange, accessed May 16, 2018, http://www.d4d.orange.com/en/Accueil.

19 "Data Collaboratives Explorer," Data Collaboratives, accessed May 16, 2018, http://datacollaboratives.org/explorer.html?#prizes-challenges.

20 AudioSet search engine can be found at https://research.google.com/audioset/.

21 Ivan Krasin and Tom Duerig, "Introducing the Open Images Dataset," *Google AI Blog*, September 30, 2016, https://research.googleblog.com/2016/09/introducing-open-images-dataset.html.

22 See Twitter's API at "Sample realtime Tweets," Twitter," accessed May 16, 2018, https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample.

23 Managing risk to affected populations is vital to any data sharing process. Protecting privacy and personal information requires detailed knowledge of the information shared and any potential risks therein. Previous research has shown that risks associated with privately collected data may be disproportionately borne by highly vulnerable populations (Madden et al. 2017). Further, any breach in data security creates risk for both the data provider and the people whose data they collect and could also undermine future data philanthropy efforts by making data providers and the general public more resistant to data sharing (Raymond et al. 2016).

24 Alina Comoreanu, "Market Share by Credit Card Network," WalletHub, March 7, 2017, https://wallethub.com/edu/market-share-by-credit-card-network/25531/.

25 Although the data to which access was provided by the Center for this use case was collected through the end of 2016, Mastercard's own internal data are constantly collected and updated in real time.

26 Information presented in this use case does not compare dollar totals, but proportions of indexed total contributions.

27 All data related to charitable giving received by Urban were aggregated by organization type according to the contribution recipient's National Taxonomy of Exempt Entities (NTEE) code. All public charities operating in the United States are assigned an NTEE code based on their primary purpose: the Mastercard team used the Urban Institute's National Center for Charitable Statistics schema to classify all charitable contributions into one of nine types before transmission to Urban, thereby prohibiting any disaggregation of results to particular charities or individuals. Readers should note, however, that these sector definitions differ from those used by Mastercard for its own *Donations Insights* reports, which classify nonprofits into 10 different categories. Further, although the category naming conventions employed in this brief overlap with those used by the *Donation Insights* reports, the actual nonprofits captured within each category may differ. Consequently, the information presented by

nonprofit type in this report should not be directly compared to the findings captured by the *Donation Insights* reports (Mastercard Center for Inclusive Growth 2016, 2017a, 2017b).

28 The education and health subsectors are known to rely more heavily on fees for services as a proportion of total revenue than other subsectors. See McKeever, Dietz, and Fyffe (2016) for estimated revenue sources by subsector.

29 2015 for NCCS data, as the 2015 NCCS Core File is the most recent year available as of publication.

30 Readers should note that Mastercard's *Donation Insights* report (Mastercard Center for Inclusive Growth 2017a) has environmental causes as the lowest average donation size. See note 27, above.

31 Timothy Sandoval, "Giving Tuesday Raises $274 Million in Its 6th Year," *The Chronicle of Philanthropy*, November 30, 2017, https://www.philanthropy.com/article/Giving-Tuesday-Raises-274/241931.

32 Timothy Sandoval, "As Giving Tuesday Expands, Nonprofits' Campaigns Grow More Sophisticated," *The Chronicle of Philanthropy*, November 7, 2017, https://www.philanthropy.com/article/Nonprofits-See-Results-as-They/241694; Nicole McGougan, "2017 Giving Tuesday Highest on Record," press release, Blackbaud, November 29, 2017, https://blackbaudnews.com/press-release/giving-tuesday-2017-donations.htm.

33 The Center's *Donation Insights* reports provide information on how giving has been affected by political events and natural disasters. For information on the share of US donations to charitable versus political organizations during presidential elections, see Mastercard Center for Inclusive Growth (2017a). For information on donations to international and foreign affairs organizations in the wake of recent natural disasters, see Mastercard Center for Inclusive Growth (2016).

34 Mychal Cohen and Peter A. Tatian, "Can a community land trust give long-term residents a foothold in a changing neighborhood?" *Urban Wire* (blog), Urban Institute, January 29, 2018, https://www.urban.org/urban-wire/can-community-land-trust-give-long-term-residents-foothold-changing-neighborhood.

35 An initial version of this framework was presented in Bogle, Diby, and Burnstein (2016).

36 For cardholder locations, east of the river was defined as zip codes 20019, 20020, and 20032. West of the river was defined as zip codes 20003 and 20024.

37 "Data For Good Exchange 2018," Bloomberg, accessed May 16th, 2018, https://www.bloomberg.com/company/d4gx/

38 Shamina Singh, "A Call to Action on Data Philanthropy," Mastercard Center for Inclusive Growth, October 4, 2016, https://mastercardcenter.org/action/call-action-data-philanthropy/

References

Bogle, Mary, Somala Diby, and Eric Burnstein. 2016. *Equitable Development Planning and Urban Park Space: Early Insights from DC's 11th Street Bridge Park Project*. Washington, DC: Urban Institute.

Butler, Declan. 2013. "When Google Got Flu Wrong." *Nature* 494:155–56.

Czajka, John L., and Amy Beyler. 2016. *Declining Response Rates in Federal Surveys: Trends and Implications*. Washington, DC: Mathematica Policy Research.

Future of Privacy Forum. 2017. *Understanding Corporate Data Sharing Decisions: Practices, Challenges, and Opportunities for Sharing Corporate Data with Researchers*. Washington, DC: Future of Privacy Forum.

Greene, Solomon, and Kathryn L. S. Petit. 2016. "What if Cities Used Data to Drive Inclusive Neighborhood Change?" Washington, DC: Urban Institute.

Groves, Robert M., and Adam Neufeld. 2017. *Accelerating the Sharing of Data across Sectors to Advance the Common Good*. Washington, DC: Beeck Center.

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

Larrimore, Jeff, Alex Durante, Christina Park, and Anna Tranfaglia. 2017. *Report on the Economic Well-Being of U.S. Households in 2016*. Washington, DC: Board of Governors of the Federal Reserve System.

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

Madden, Mary, Michele Gilman, Karen Levy, and Alice Marwick. 2017. "Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans." *Washington University Law* Review 95 (1): 53–125.

Mastercard Center for Inclusive Growth. 2016. *Donation Insights: US Report for the Period Ended September 2016*.

———. 2017a. *Donation Insights: US Report March 2017*.

———. 2017b. *Donation Insights: Closing the Information Gap on Charitable Giving: US Report October 2017*.

McKeever, Brice, Nathan Dietz, and Saunji Fyffe. 2016. *The Nonprofit Almanac: The Essential Facts and Figures for Managers, Researchers, and Volunteers,* 9th ed. Washington, DC: Urban Institute.

Raymond, Nathaniel, Ziad Al Achkar, Stefaan Verhulst, and Jos Berens. 2016. "Building Data Responsibility into Humanitarian Action." New York: United Nations Office for the Coordination of Humanitarian Affairs.

Verhulst, Stefaan, and Andrew Young. 2016. *Open Data Impact: When Demand and Supply Meet: Key Findings of the Open Data Impact Case Studies*. New York: GovLab.

———. 2017. *The Potential of Social Media Intelligence to Improve People's Lives: Social Media Data for Good*. New York: GovLab.

Wesolowski, Amy, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. 2012. "Quantifying the Impact of Human Mobility on Malaria." *Science* 338 (6104): 267–70.

About the Authors

Brice McKeever is a research associate in the Center on Nonprofits and Philanthropy at the Urban Institute, where he performs quantitative social science research across a wide spectrum of public policy initiatives related to the nonprofit sector, civic participation, and the impact of technology on social interaction. He is the lead author of *The Nonprofit Almanac: Ninth Edition* and multiple research reports related to nonprofit-sector growth and diversity.

Solomon Greene is a senior fellow in the Research to Action Lab and the Metropolitan Housing and Communities Policy Center at the Urban Institute. His research focuses on how land-use and housing policy can improve access to opportunity for vulnerable groups and how data and technology can support more inclusive urban development.

Graham MacDonald is chief data scientist at the Urban Institute, where he works with researchers to improve access to data, analytics tools, and innovative research methods. MacDonald uses and advises on such tools as machine learning, natural language processing, web scraping, big data platforms, and data visualization techniques and their application to relevant public policy issues.

Peter Tatian is a senior fellow at the Urban Institute and research director for Urban–Greater DC. He advises nonprofits on performance management and evaluation and heads Urban's work providing technical assistance on data collection and use to grantees of the US Department of Education's Promise Neighborhoods initiative.

Deondre' Jones is a research assistant in the Center on Nonprofits and Philanthropy, where he supports research, writing, and analysis on projects related to program evaluation and outcome measurement. Before joining Urban, Jones served in the 23rd class of the Emerson National Hunger Fellowship, a program hosted through the Congressional Hunger Center. He worked on food security and national antipoverty policy issues.

URBAN INSTITUTE · ELEVATE · THE · DEBATE

2100 M Street NW
Washington, DC 20037

www.urban.org