ie UNIVERSITY
CENTER FOR THE GOVERNANCE OF CHANGE

JULY 2023

DATA COLLABORATIVES: ENABLING A HEALTHY DATA ECONOMY THROUGH PARTNERSHIPS

cgc.ie.edu

CONTENTS

AUTHOR:
Dr. Stefaan G Verhulst

INTRODUCTION

When did our current era begin? One plausible start date is September 9, 2016. That's when the total amount of Internet traffic exceeded one zettabyte—officially inaugurating what some have called the Zettabyte Era (or, alternatively, the Zettabyte Zone).

The scale boggles the mind, and is a testament to the rapid datafication of our society. A Zettabyte is 10 to the power of 21 bytes—one trillion gigabytes. If the gigabytes in a zettabyte were broken down into meters, then one zettabyte would cover 150,000 times the distance of the Amazon. If a gigabyte were a brick, then a zettabyte would be equivalent to 258 Great Walls of China (3,873,000,000 bricks).[1]

> The datafication of virtually every aspect of our private and public lives presents both opportunities and challenges. Among the most important of these challenges is the emerging problem of data asymmetries—the uncomfortable reality presented by scarcity amid a time of unprecedented plenty. Although our society is awash in data, it is increasingly clear that data and its benefits are not equally distributed. Instead, data flows have grafted themselves onto existing, and deeply entrenched, inequalities in our society, in many cases, exacerbating them.

Overcoming data silos is key to addressing these data asymmetries and promoting a healthy data economy. This is equally true of silos that exist within sectors as it is of those among sectors (e.g., between the public and private sectors). Today, there is a critical mismatch between data supply and demand. The data that could be most useful rarely gets applied to the social, economic, cultural, and political problems it could help solve.

> Data silos, driven in large part by deeply entrenched asymmetries and a growing sense of "ownership," are stunting the public good potential of data.

This paper presents a framework for responsible data sharing and reuse that could increase sharing between the public and private sectors to address some of the most entrenched asymmetries. Drawing on theoretical and empirical material, we begin by outlining how a period of rapid datafication—the Era of the Zettabyte—has led to data asymmetries that are increasingly deleterious to the public good. Sections II and III are normative. Having outlined the nature and scope of the problem, we present a number of steps and recommendations that could help overcome or mitigate data asymmetries. In particular, we focus on one institutional structure that has proven particularly promising: data collaboratives, an emerging model for data sharing between sectors. We show how data collaboratives could ease the flow of data between the public and private sectors, helping break down silos and ease asymmetries. Section II offers a conceptual overview of data collaboratives, while Section III provides an approach to operationalizing data collaboratives. It presents a number of specific mechanisms to build a trusted sharing ecology.

I.
THE CASE FOR COLLABORATION: BIG DATA, DATAFICATION, AND DATA ASYMMETRIES

I. THE CASE FOR COLLABORATION: BIG DATA, DATAFICATION, AND DATA ASYMMETRIES

This section seeks to make a case for why more—and better—collaboration is necessary to address data asymmetries across society. Beginning with a general overview of the process of datafication, it argues that an era of plenty is, paradoxically, also marked by scarcity, silos, and asymmetries. These challenges are pervasive and may be exacerbating. They draw attention to the urgent need for more sharing through data collaboratives and other mechanisms—which we explore in Section II.

A. BIG DATA

To understand the Zettabyte Era, it is useful to begin with the concept of big data. The term has in recent years gained increasing currency as a way of describing a cross-sectoral phenomenon resulting from widespread digitalization. Typically, it is understood to refer to a quantitative phenomenon—i.e., characterized by the proliferation or abundance of data. However, big data extends beyond mere *bigness*; understanding its related properties can also help us understand the phenomenon of datafication.

Many competing definitions of big data exist, but there is widespread agreement that it cannot simply be defined by size or volume.[2] Mike Loukides, for example, argues that the "big" in "Big Data" is a "red herring." He points out that both the public and private sector have long handled large datasets and argues that "Big Data" must be understood as occurring when the size (or amount) of data itself becomes part of the problem."[3] A recent editorial in *Nature* magazine likewise argues that "'Big'… is a moving target";[4] it points out that amounts of data that would have seemed huge a few years ago are now routinely carried around on portable pen drives.[5]

The volume of data is undeniably part of what makes big data big. Nonetheless, the following characteristics must also be considered:

- **Three Vs:** Big data can be characterized by three Vs—Velocity, Variety, and Veracity.[6] Velocity refers to the speed and constantly accelerating the rate at which data is produced. Variety refers to the different forms in which data is being generated and, in particular, the sheer variety of unstructured data, which represents a host of analytical and other processing challenges (see below). Veracity refers to the problem of accuracy or—in more technical terms—the quality that is inherent to big data.
- **Unstructured Information:** One of the key characteristics of the era of big data is the proliferation of and ability to find meaning in unstructured data. According to one estimate, only around 5% of information created today is "structured," which is defined as information that "comes in a standard format of words or numbers that can be read by computers."[7] The vast majority of data exists in the form of phone calls, notes, photos, and other non-standardized formats. The era of big data is marked by algorithmic advances that make it easier to work with unstructured information, as well as crowd-sourced and open-source tools that permit easier detection of signals within noise.[8]
- **Commingled Data:** Big data is also marked by the commingling or aggregation of disparate large databases to extract new relations and patterns. The ability to combine information from different sources and extract meaning from it is sometimes referred to as a process of turning "dross into gold"—without modern data analysis tools, much of the unstructured data would simply go unused.[9] This fundamental "relationality" of big data offers tremendous promise to a wide variety of fields and sectors, both public and private.[10]

"

The concept of datafication is often discussed primarily as a commercial phenomenon, and its value as such is undeniable. It is important to recognize, however, that the value of datafication extends far beyond what's simply monetizable—and this has important ramifications for the era in which we are living.

B. DATAFICATION

The process of datafication emerges directly from the phenomenon of big data. Datafication can be said to exist on a foundation of big data. In this sense, the traits outlined above are critical to the notion of datafication, yet they do not capture the full phenomenon.

> Understanding some of the unique drivers and characteristics of datafication can help us better understand some of the resulting asymmetries and therefore the need for more sharing.

i. Drivers of datafication: The emergence of datafication has been enabled by numerous factors, including:

- Changes in the way data is collected, including a proliferation of digital sensors and personal digital devices, resulting in ever-widening streams of "digital exhaust" or "data exhaust";[11]
- Changes in the way data is stored, including the rise of cloud computing and (virtually) unlimited memory;
- Changes in computation and analytic capacities, driven by advances in computational and data science, and the rise of artificial intelligence, machine learning, and new methods of data visualization;
- Changes in the use of and reliance on data and data insights, by businesses and the public sector and the accompanying rise of evidence-based decision-making.

ii. Characteristics of Datafication: The concept of datafication is often discussed primarily as a commercial phenomenon, and its value as such is undeniable. It is important to recognize, however, that the value of datafication extends far beyond what's simply monetizable—and this has important ramifications for the era in which we are living. As Mejias and Couldry argue in the *Internet Policy Review*, datafication has also resulted in "the transformation of human life into data through processes of quantification," and this transformation, the authors further argue, has "major social consequences … [for] disciplines such as political economy, critical data studies, software studies, legal theory, and—more recently—decolonial theory." In this sense, datafication can be understood as a fundamentally social, cultural and sociological phenomenon.

Three key features of datafication are worth highlighting; they help us understand how data plenty has led to deeply entrenched asymmetries, and why more data sharing is essential.

- Datafication is *all pervasive*, which means it permeates and emanates from virtually every aspect of citizens' lives. Sometimes referred to as a process of "life mining,"[12] datafication emerges from the data trails left behind by citizens' use of social media, sensors and personal devices like telephones and GPS equipment, as well as various other nodes on the Internet of Things (IoT).
- The resulting "exhaust trails" are, as a result, deeply *socially contextualized*. Reflecting virtually our entire social lives, they by extension contain our social, economic, and political patterns. Datafication therefore effectively involves digitalizing and building a quantifiable map of social exclusion. As Mejias and Couldry argue, many analyses of datafication explain its nature and significance "in terms of its relationship to time, context, and power."[13]
- All of this in effect means that our data ecology is today profoundly *reflective of our social asymmetries*. Like much technology and science in general, data is often normalized and presented as neutral. As scholars

have argued, however, data results in "nothing less than a new social order, based on continuous tracking, and offering unprecedented new opportunities for social discrimination and behavioral influence."[14] Furthermore, data does not simply contain an imprint of existing hierarchies and inequalities; it also perpetuates them. These asymmetries and patterns of exclusion explain the importance of breaking down data silos and increasing data sharing.

C. DATA ASYMMETRIES

Much attention has been paid in recent years to the challenges (or negative externalities) associated with datafication.[15] The problems commonly highlighted include those related to "dataveillance, "[16] the emergence of "surveillance capitalism,[17] and data extraction without consent.[18] In addition, scholars have written about the risks of "data colonialism"[19] and threats to individual autonomy and dignity.[20] As we note above, however, data asymmetries stand out as among the most critical of externalities.

Data asymmetries often result from data hoarding or the "industrial complex"[21] that exists behind datafication. They occur whenever there exists a divide or disparity in access to and re-use of data.[22] The nature of this divide can take many forms, depending on the relationship between data holders, data subjects and users. A non-exhaustive list of data asymmetries is included in the Appendix. It includes imbalances between citizens and private sector or government entities, as well as those between the public and private sectors.

> Each of these manifestations poses unique challenges and problems. Considered together, however, they make clear the wider stakes, and suggest the urgent need for more sharing within and among sectors.

Mejias and Couldry argue that "fundamental to [an understanding of datafication] is the analysis of the intersection of power and knowledge."[23] As we have elsewhere argued, many of our society's patterns of exclusion and inequalities are therefore refracted through patterns of access in the wider data ecology. If, as scholars like Thomas Piketty and others have argued,[24] overcoming inequalities is the defining challenge of our era, then inequalities within the data ecology represent a particularly troublesome aspect of that challenge in its ability to enable or otherwise perpetuate other inequalities. In the following sections, we explain how greater sharing can help ease these inequalities, and make the case for a particular mechanism for sharing: data collaboratives.

II.
TOWARD SOLUTIONS: DATA SHARING AND THE POTENTIAL OF DATA COLLABORATIVES

II. TOWARD SOLUTIONS: DATA SAHRING AND THE POTENTIAL OF DATA COLLABORATIVES

> The pathway to more data sharing runs through the well-established (yet still poorly understood) practice of open data.

Earlier efforts at opening data, beginning several decades ago, have made considerable progress in easing data silos and the resulting asymmetries related to government-derived data. Yet they have also fallen short in important ways. As we have elsewhere argued, we are now entering a "third wave of open data."[25] This third wave offers particular potential for data reuse and sharing.

A. A THIRD WAVE OF OPEN DATA

i. The First Two Waves: The first wave of open data was marked by the institution of Freedom of Information (FOI) and other related laws, which made national government data available on request to an audience (largely) composed of journalists, lawyers, and activists. This era dated roughly from the late 1990s. The second wave of open data was enabled by the advent of open source and Web 2.0. It called upon governments to make their data open by default (rather than request) to civic technologists, government agencies, and corporations.

These first two waves achieved many successes, not least of which were establishing the key principles that data should be open and that its reuse could result in social benefits. Yet the results of these efforts were also incomplete. For example, much of the released data focused on national and supra-national organizations even though larger amounts of data were held in silos at the subnational, local level. As a result, many data-associated asymmetries have persisted, especially regionally. Today, much—perhaps even a majority—of generated data remains locked away and inaccessible to those who need it most.

ii. Third Wave: The third wave of open data seeks to build on earlier successes to further ease data asymmetries. It does this by focusing on the demand for data as much as the supply, seeking to understand the broader technical, social, political and economic context within which data is produced and consumed. Based on original research conducted by the author, the third wave includes the following characteristics and goals:

- ***Publishing with Purpose,*** and in particular an attempt to better match data supply with demand;

- ***Fostering Partnerships and Data Collaboration***, between the public and private sectors in particular, but also more generally across and within sectors;
- ***Easing Data Asymmetries at the Subnational Level***, for instance by providing resources to cities, municipalities, states, and provinces to better access and share information.
- ***Prioritizing Data Responsibility and Data Rights***, by understanding the risks of using (as well as not using) data in service of the public good.[26]

B. POTENTIAL OF THE THIRD WAVE

At a high level, the potential of this Third Wave rests in its ability to break down data silos and facilitate sharing among sectors. This holds tremendous possibilities for positive social transformation, much of which would be mediated by the resulting flattening of hierarchies and asymmetries. More open data means more access to data or data products (at least in theory). It means that the potential insights contained within data can be better directed in service of those who may most benefit from those insights, as well as those who may be in a best position to unlock the insights. This may be particularly true of sharing between the public and private sectors. As we argued in a 2019 paper titled "Leveraging Private Data for Public Good:"

> "[M]uch of the most useful, timely and comprehensive data that could help transform the way we make decisions or solve public problems resides with the private sector in the form of call detail records, online purchases, sensor data, social media data, and other assets. If we truly want to harness the potential of data to improve people's lives, we need to understand and find ways to unlock and re-use this private data for public good."[27]

In parallel to unlocking private sector data, it is also important for the public sector to increase the availability of data for the private sector and society at large. Recognizing the immense value that data holds for innovation and societal benefits, the public sector should actively work towards fostering a culture of data sharing and openness. By making relevant data sets more accessible, policymakers can unlock opportunities for the private sector to develop innovative solutions, drive economic growth, and improve overall quality of life.

In addition to these high-level benefits, there are four further (associated) value propositions for more data sharing:

i. Situational Awareness: First, data sharing can help improve situational awareness for both private and public sector entities, as well as the effectiveness and speed of responses to crises. For instance, data held by the private sector can help government agencies better understand demographic trends, public sentiment, and the geographic distribution of phenomena, such as pandemics or illnesses, in the process designing better responses.[28]

A good example can be found in Chile, where telecom data accessed by the Gender and Urban Mobility data collaborative—an organization that includes various stakeholders including The GovLab, UNICEF, Universidad del Desarrollo, Telefónica R&D Center, ISI Foundation, and DigitalGlobe—yielded a better understanding of the mobility experiences of women and girls in Santiago de Chile. The insights provided urban planners with a greater understanding of gendered differences in how residents move around the city. This allowed planners and other agencies to be more sensitive in their designs to the needs of girls and women.[29]

ii. Understanding Cause and Effect: Datasets shared across sectors can also be combined and analyzed to better link cause and effect, in the process ensuring that those responsible for solving problems have greater insight into the phenomena driving crises and other social ills.

> Recognizing the immense value that data holds for innovation and societal benefits, the public sector should actively work towards fostering a culture of data sharing and openness.

A notable example of a data initiative that sought to identify cause and effect was the use of open government and private-sector data in the response to the Ebola outbreak of 2014, which brought together different datasets generated by governments, international organizations, humanitarian responders and telecom carriers, among others, to help responders identify how and why the virus spread.[30]

iii. Prediction and Forecasting: Third, more sharing can create new predictive capabilities through the analysis and combination of previously inaccessible datasets. Such data-driven forecasting capabilities can help institutions and policymakers be more proactive and avert crises before they occur.

For example, in 2019, an Indian effort to leverage Bing data offered insight into potential drivers of suicide ideation among teenagers in India, helping public-sector actors to improve preventative outreach and interventions to those exhibiting behavior that could lead to self-harm (e.g., by identifying searches for certain toxins that could be used for suicide).[31]

iv. Monitoring and Evaluation: Finally, more data sharing can help institutions monitor and evaluate the impact of policies and interventions, often in real-time. This helps government agencies design better, more responsive, and evidence-based policies or services, and enables a process of iteration and constant experimentation.

A good example can be found in the US Food and Drug Administration, which established the Sentinel Initiative (sentinelinitiative.org). The initiative uses a distributed database to run analytical programs on local databases of private-sector health providers (such as Humana Inc. and Blue Cross Blue Shield), allowing the agency to securely analyze safety data in order to monitor adverse reactions to medical products on the market.

C. DATA COLLABORATIVES

Despite the clear benefits of data sharing, backed up by a growing body of evidence, too much data remains in silos. There are many reasons for this bottleneck, including a search for competitive advantage, regulatory caution, and general distrust of sharing and data reuse. To an extent, the overarching problem remains a paucity of credible models.

In recent years, one model has gained new valence, and has been used with increasing frequency by both public and private sector entities: data collaboration. Much of our work has focused on the potential of this mechanism. In the remainder of this paper, we focus on data collaboratives: their potential, their challenges, and pathways to implementation.

i. What are Data Collaboratives?

The term data collaborative refers to an emerging model of collaboration in which participants from different sectors—including private companies, research institutions, and government agencies—exchange data to help solve public problems.[32] Data collaboratives are key to overcoming many of the bottlenecks and asymmetries within the data ecology. While much commentary is today focused on the glut of available data, in fact, as we have noted, data supply and demand are often poorly matched: those who most need data, or who could most productively use it, often don't have access to it. Thus one of the key challenges of our era lies in a persistent failure to reuse data responsibly for public good. This failure results in tremendous inefficiencies and lost potential. Data collaboratives address these shortcomings by drawing together otherwise siloed data and a dispersed range of expertise, matching supply and demand, and ensuring that relevant institutions and individuals are using and analyzing data in ways that maximize the possibility of new, innovative social solutions.

We coined the term "data collaborative" in 2015.[33] By 2018, we had identified 145 data collaboratives in our repository from across the world in 11 different sectors, testifying to the potential and realized contributions of this mechanism.[34] This has recently been updated to more than 200 examples.

ii. Models for Data Collaboratives

As we move from theory to a practice of data collaboratives, certain patterns are becoming clearer. Data collaboratives are not a uniform phenomenon. Especially as they spread around the world and sectors, we are seeing variations emerging. It is important to consider these patterns and variations in order to better understand what works (and what doesn't) when it comes to data sharing.

In our research, we observe six different types of data collaboratives, each offering their own lessons (and cautions) for the goal of data sharing:

- ***Public Interfaces:*** Companies provide open access to certain data assets, enabling independent uses of the data by external parties. Current approaches include: APIs and Data Platforms.
- ***Trusted Intermediary:*** Third-party actors support collaboration between private-sector data providers and data users from the public sector, civil society, or academia. Current approaches include: Data Brokerage and Third Party Analytics Projects.
- ***Data Pooling:*** Companies and other data holders agree to create a unified presentation of datasets as a collection accessible by multiple parties. Current approaches include: Public Data Pools and Private Data Pools.
- ***Research and Analysis Partnerships:*** Companies engage directly with public-sector partners and share certain proprietary data assets to generate new knowledge with public value. Current approaches include: Data Transfers and Data Fellowships.
- ***Prizes and Challenges:*** Companies make data available to participants who compete to develop apps; answer problem statements; test hypotheses and premises; or pioneer innovative uses of data for the public interest and to provide business value. Current approaches include: Open Innovation Challenges and Selective Innovation Challenges.
- ***Intelligence Generation:*** Companies internally develop data-driven analyses, tools, and other resources, and release those insights to the broader public.

The field of data sharing is characterized by a pervasive absence of trust. This is true both among potential sharing partners and also among the public, which remains ambivalent and skeptical about how its data is being (re)used.

D. CHALLENGES OF DATA COLLABORATIVES

Data collaboratives offer a promising model for data sharing and collaboration across sectors. The benefits of greater data sharing are outlined above. Yet it is also important to keep in mind that data collaboratives—like any effort at data sharing—also pose certain risks. In order to design operational models to facilitate responsible data sharing (Section III), we need to understand both the opportunities and challenges.

Based on our research, we identify the following main challenges:

i. Lack of Awareness and Data Literacy: Both among those who hold data and those who might use it (suppliers and consumers) there often exists a lack of awareness and appreciation regarding the potential of data sharing. This can take the form of a general lack of awareness about the opportunities (and challenges) of data reuse, or it may represent a lack of understanding about a particular opportunity—i.e., a recognition of how a particular dataset can be directed to help solve a particular public challenge.

ii. Absence of Trust: The field of data sharing is characterized by a pervasive absence of trust. This is true both among potential sharing partners and also among the public, which remains ambivalent and skeptical about how its data is being (re)used. While such concerns are understandable and often valid, the absence of trust acts as a barrier to the potential of data sharing. It strongly suggests the need for a responsible data sharing framework, something we discuss further below. Such a framework could help build trust, especially if it is made publicly available, includes a fair allocation of liability and dispute resolution mechanisms, and is accompanied by robust steps for monitoring and to ensure accountability.

iii. Uncertainty within the Private Sector (Unclear Incentives): Despite clear evidence for the benefits of data sharing, companies often have concerns and reservations about the reuse of their data. Some of these concerns are no doubt legitimate, but they act as a barrier to unleashing the potential of data for the public good. A (partial) list of concerns include:

- Data leaks and competitors gaining business intelligence about markets and operations;
- Penalties and fines by regulators or other lawmakers imposed due to the interpretation of (often unclear) legislation and processes; and
- Reputational loss if customers grow suspicious of how their data is being used and recycled.

Addressing these concerns, and developing a clearer set of incentives for the private sector, is critical to enabling more data sharing.

> Despite clear evidence for the benefits of data sharing, companies often have concerns and reservations about the reuse of their data. Some of these concerns are no doubt legitimate, but they act as a barrier to unleashing the potential of data for the public good.

iv. Limited Capacity: The ability to process, analyze and use data varies widely by organization, another factor which limits sharing and the overall public good potential of data. This lack of capacity can manifest as a lack of technical knowledge (e.g., insufficient data skills), financial resources, or simply as a lack of awareness. Capacity limitations are particularly a problem for poorly funded government agencies, as well as for smaller private- and public-sector entities, which may similarly lack adequate technical and financial means to foster a sharing culture.

v. Transaction Costs: While open data is often (though not always) made available without charge, it would be incorrect to assert that data sharing is always free of cost. Transaction costs are incurred throughout the data life cycle—while preparing data; de-risking data (e.g. through anonymization); and in coordinating with partners, including through the preparation of legal agreements or other structures, mechanisms or institutions to permit data sharing and reuse. These transaction costs can inhibit an organization's willingness to share and reuse data (without a fair compensation scheme[35]).

vi. Limited Community of Practice and Knowledge Base: Finally, the nascent nature of data sharing poses an additional barrier. Successful initiatives require a community of practice and build upon an established knowledge base (including, for example, case studies and lessons learned). Although the situation is improving as data collaboratives and other mechanisms become more established, we still note an overall absence of a sharing culture to facilitate true collaboration among sectors. Over time, as data-sharing initiatives multiply, we would expect to see the emergence of new bodies, institutions, and bodies of knowledge that could offer a more solid foundation for a community of practice and learning.

III.
PATHWAYS FORWARD: MAKING DATA COLLABORATIVES SYSTEMATIC, SUSTAINABLE AND RESPONSIBLE

III. PATHWAYS FORWARD: MAKING DATA COLLABORATIVES SYSTEMATIC, SUSTAINABLE AND RESPONSIBLE

The preceding section has argued for the societal benefits of greater data sharing, and for the potential of data collaboratives. While not without their own risks and challenges, data collaboratives provide a useful model to increase sharing between the public and private sectors, in the process helping to ease some of the most persistent challenges of the data ecology as well as associated ones within society at large.

If Section II was *normative*, then this section can be considered *operational*. We examine the following concrete steps or mechanisms for making data collaboratives more systematic, sustainable and responsible. In the process, we begin fleshing out an operational framework for greater data sharing between the private and public sectors, and more generally among and within sectors.

A. A NEW SCIENCE OF QUESTIONS

The sheer variety and complexity of challenges facing our world can be overwhelming. We know that data—and shared data in particular—can be helpful in a variety of contexts. Nonetheless, policymakers and other stakeholders are presented with problems of prioritization, especially in a world of limited resources. Establishing a new science of questions can help identify the most promising public challenges that are amenable to a data fix. Such a science could also help operationalize data collaboratives by identifying what types of data should be shared, with whom, and through what mechanisms.

In 2019, the GovLab, in collaboration with Schmidt Futures, launched The 100 Questions initiative.[36] The initiative sought to establish priorities by mapping the world's 100 most pressing, high-impact questions that could be answered if relevant datasets were made available. In collaboration with global thought leaders, The GovLab developed a new participatory methodology to define and prioritize questions and societal problems.

The initiative has thus far helped identify eight problem areas (including migration, air quality, gender, the future of work) and ten questions within each of these areas.[37] As important as the specific questions and problem areas, however, was the methodology deployed to arrive at them—i.e., the elements of a new science of questions.

Two of the most important elements of this new science were a *smarter crowdsourcing* method, and a *reliance on a cohort of "bilinguals."* As a methodology, the former sought to attract diverse ideas from global experts, thus combining the reach and openness of traditional crowdsourcing, which seeks input from the public, with the rigor of expert opinion and scientific method. By bilinguals, we mean experts who possess both domain specific knowledge (i.e., relevant to a particular problem area) and those who are data science specialists. This ensures not only that the identified problems are relevant and important but also that they are amenable to a data solution. Each of these elements, we believe, can contribute to a more effective, and responsible data sharing framework.

B. PROFESSION OF DATA STEWARDS

Effective sharing relies on various factors. One of the most important factors is whether there exists within sharing organizations individuals or teams specifically empowered to proactively initiate, facilitate and coordinate data collaboratives.[38] We call such individuals and teams "data stewards." They are a critical part of the "open data stack" within organizations around the world, and essential to fostering a culture of responsible sharing.[39]

Data stewards have the requisite expertise and authority to recognize opportunities for productive collaborations or to respond to external requests for data. They systematize the process of partnering, and help scale efforts when there are fledgling signs of success. More specifically, data stewards perform the following five functions:

PARTNERSHIP AND COMMUNITY ENGAGEMENT:

Data stewards develop and implement a more proactive and responsive approach to reaching out to and vetting potential sharing partners. They should be informing potential beneficiaries (and others) of the possibilities of data collaboration, engaging with all actors (both within and without their organizations) who may be affected by sharing, and generally fostering a sharing culture within both the private and public sectors.

INTERNAL COORDINATION AND STAFF ENGAGEMENT:

Establishing a successful data collaborative requires internal coordination and sign-off from various stakeholders—including, for instance, legal, policy, technical, data, marketing and sales teams. Data stewards are key to ensuring internal stakeholders and company leadership are informed and aligned. In addition, data stewards often play an important role in mapping and matching staff with specific skills, such as data science abilities, or interest in data collaborative initiatives.

DATA AUDIT, ETHICS AND ASSESSMENT OF VALUE AND RISK:

Data stewards are responsible for monitoring and assessing the value, potential, and risk of all data held within organizations. This responsibility includes knowing what data an organization collects, and what public interest questions that data could potentially help answer if shared.

DISSEMINATION AND COMMUNICATION OF FINDINGS:

Data stewards often act as the public face of an organization's data projects, and they are responsible for raising awareness, disseminating findings and communicating shared outcomes from data collaboratives. Data stewards may also be responsible for overall communication with customers, citizens, partners, and other stakeholders about regulatory compliance, contractual obligations and how data is being shared and used, and what public benefits it may have.

NURTURE DATA COLLABORATIVES TO SUSTAINABILITY:

Many ambitious data collaborative projects collapse after initial pilots or experiments. Data collaboratives can play a valuable role (partly through their dissemination and communication functions) in nurturing and helping scale these projects until they are sustainable. While data stewards may not themselves have the requisite budget to ensure long-term sustainability themselves, they must work with a variety of stakeholders to gather the needed resources and support so as to ensure broad and long-term impact.

DECARBONIZATION:

More recently a new area of responsibility has emerged: the decarbonization of data. The rapid expansion of digital technologies and the exponential increase in data generation have contributed to a significant environmental impact, primarily through energy consumption and carbon emissions. Data centers, network infrastructure, and the computational power required for data processing contribute to greenhouse gas emissions and energy consumption on a substantial scale. Recognizing this, organizations and data stewards are increasingly focusing on mitigating the carbon footprint associated with data activities. This involves adopting sustainable practices, such as optimizing energy efficiency, transitioning to renewable energy sources, and designing data infrastructure that minimizes environmental impact. Furthermore, data stewards are exploring innovative internal solutions like data compression techniques, data deduplication, and data lifecycle management to reduce storage and processing requirements, thereby minimizing energy consumption. Decarbonization of data is not only an ethical imperative but also a strategic move toward building a more sustainable future. By integrating environmental considerations into data stewardship practices, organizations can effectively balance the benefits of data-driven technologies with the imperative of mitigating climate change and promoting ecological sustainability.

C. CLARIFY INCENTIVES

The case for sharing cannot rest on altruism alone, yet a lack of clarity surrounding incentives is one of the major impediments to greater data collaboration. In fact, concerns over perceived competitive threats or regulatory repercussions (e.g., for sharing PII) can actively disincentivize sharing, especially by the private sector. Actively making the case for the benefits of sharing is thus key to operationalizing more data collaboration.

In a 2021 report, this author, along with collaborators, proposed the "9Rs Framework"[40] to clarify incentives. Although primarily directed at sharing by the private sector, many of its components are also relevant to the public sector:

1. **Reciprocity:** Gaining access to data sources and other assets held by organizations whose data may be important to business decisions and result in competitive advantage;

2. **Rectifying Errors and Improving Data Quality:** Identifying errors in datasets by letting others access, analyze and use them;

3. **Research and Insights**: Generating new answers to questions, and providing organizations with insights that may not have otherwise been extracted;

4. **Reproducibility:** Testing results of analysis by allowing others to conduct identical or related work;

5. **Reputation**: Enhancing an organization's image and reputation, attracting media, new users, customers, and investors who value socially conscious corporate actors;

6. **Responsibility and Philanthropy:** Fulfilling an organization's social responsibilities, improving the environment in which it operates, and bolstering its reputation;

7. **Recruiting and Retaining Talent:** Attracting and retaining data science talent with projects that are compelling and socially relevant;

8. **Regulatory Compliance:** Helping organizations comply with regulations, become more transparent, or otherwise promote responsible data management; and

9. **Revenue Maximization:** Providing opportunities to generate new income or cut costs.[41]

D. ESTABLISH A SOCIAL LICENSE FOR RE-USE

Clarifying incentives can help make the business (and financial) case within organizations for data sharing. But as much as operationalizing collaboration depends on incentivizing data holders, its success ultimately rests on making the case more broadly—to various stakeholders in society at large. We call this establishing a *social license* for data collaboration.

Trust is a foundational principle of any social license. If data collaboration is to be operationalized at a scale that can have genuine transformative impact, then all stakeholders—data holders, data consumers, and most importantly the public at large—must be able to trust that all parties will uphold their responsibilities when it comes to how data is collected, stored, and used. Transparency, open dialogue, and social contracts can help to build trust between involved parties. Explainability and data literacy are also key concepts, as without these, information asymmetries between the public and data practitioners can create power differentials, thus weakening trust.[42]

Based on our research, we conclude that three approaches are particularly important in securing trust:

1. **Public engagement:** Since any social license depends on the public's approval—or at least acceptance—public engagement acts as the foundation of social licenses. Public engagement can take many forms, from data literacy campaigns that build community awareness, to citizen assemblies, to open dialogue between stakeholders in order to foster better mutual understanding.[43] Public engagement is important not only to establish trust and social contracts between the public and data practitioners, but also to create opportunities for honest assessments of the benefits and risks associated with any given project.[44]

2. **Data stewardship:** The notion and vital function of data stewardship are outlined above. Data stewards play a key role in securing a social license for data reuse, for instance by leveraging their role as conduits between various stakeholders and data practitioners to establish a rapport of trust and open communication. They are also well positioned to implement principles like transparency and explainability, as well as ensure the adoption of responsible data practices that make it easier to build trust.[45]

3. **Regulatory framework:** An enabling regulatory framework is also critical to operationalizing data sharing by building trust. Social licenses have sometimes been conceived of as bridges between what is legally permissible and what is socially acceptable.[45] In this conception, regulatory frameworks help define legal limits, thus establishing a baseline for social licenses. Furthermore, a strong regulatory framework may give the public greater confidence in actors involved in data sharing, helping to foster trust. Most importantly, regulatory frameworks create a system of accountability wherein breaches in the terms of a social license by data practitioners can be addressed and action taken. This gives the public a stronger position to negotiate from and empowers stakeholders to more concretely implement their social licenses.

> Although our world is today awash in data, a persistent paucity of data continues to stunt the potential of data collaboration.

E. BECOMING DATA-DRIVEN ABOUT DATA

Our final operational recommendation derives from an often-unacknowledged paradox at the heart of the data sharing ecology: Although our world is today awash in data, a persistent *paucity* of data continues to stunt the potential of data collaboration. For all the examples that suggest the power of data sharing, the evidence base remains thin, and both the theory and practice of collaboration are unsystematized and under-researched.

This doesn't just hamper our understanding of data sharing. It also limits the reproducibility and scalability of data projects. Without more knowledge of what works (and what doesn't), it is harder to establish a database of best practices or operational guidelines, such as those outlined here, in order to build sustainable and responsible data sharing initiatives. Similarly, without a better understanding of impact, it is harder to iterate to improve initiatives or achieve accountability for projects that cause direct or indirect harms.

Becoming more data-driven about data is thus a critical step in operationalizing a collaborative ecology.[47] In particular, we need more data on:

- What are the kinds of projects that organizations seek data to address?
- What types of data are being shared, and how?
- Who is sharing data, and with whom? In which domains are data collaboratives most common?
- How is that data sharing taking place and through what mechanisms? What was successful and what failed? And why?
- What safeguards are being implemented to ensure adequate protections of shared data—and, importantly, which safeguards are most effective?
- What is the impact—either positive or negative—of data sharing initiatives?
- What is the specific role of data collaboratives and data stewards (two emergent institutional structures in the sharing ecology) in sharing initiatives? In what ways are they playing an enabling vs. inhibiting role?
- How long are data collaboratives maintained and under what conditions are they ended?

IV. CONCLUSION

IV. CONCLUSION

IV. CONCLUSION

In conclusion, data collaboratives offer a promising solution to address data asymmetries in our society, but they require a systematic, sustainable, and responsible approach to be successful. A new science of questions can help identify the most pressing public and private challenges that can be addressed with data sharing. Data stewards are essential to fostering a culture of responsible sharing within organizations, and clarifying incentives are crucial to operationalizing data collaboration. Additionally, building a social license for data reuse through public engagement, data stewardship, and an enabling regulatory framework is key to establishing trust between all stakeholders involved. Finally, becoming more data-driven about data is essential to improving our understanding of collaboration, building sustainable initiatives, and achieving accountability for projects. By being smart about incentives and adopting a responsible and sustainable approach, data collaboratives can contribute to a health data economy that benefits society as a whole.

APPENDIX

APPENDIX: TYPOLOGY OF DATA ASYMMETRIES

- **Data asymmetries between citizens and corporate (B2C) or governmental (G2C) actors.** Such asymmetries have grown increasingly common with the datafication of consumption patterns or government engagements and typically occur when organizations collect data on their users while providing services or selling goods (often called two-sided markets). Typically, companies and governments often possess a disproportionate amount of data on their users and citizens—information that users may not even be aware of having surrendered.
- Other include **business-to-business (B2B) data asymmetries.** Recent years have witnessed the emergence of several large data monopolies that dominate their sectors and the broader economy. These companies have access to huge amounts of data collected and processed across various domains (such as search data, location and mobile phone data, consumer spending data) and their ability to combine and derive insights from this data or train ML algorithms results in de facto barriers to entry. There are concerns that B2B data asymmetries may be stifling innovation and competition as well as hurting the rights of consumers, leading to calls for greater regulation and better enforcement of antitrust law, perhaps extending so far as to the breakup of some of these large players.
- Policymakers are increasingly turning their focus to **data asymmetries in the Business-to-Government space.** Multi-sectoral actors, including the High-Level Expert Group to the European Commission on B2G Data Sharing,[48] of which I was a member, are shining a light on the ways in which government decision-making and service delivery can be hampered by the lack of access to data and insights that are held in the private sector and used for commercial purposes.
- In addition, the field has paid comparatively little attention to another important data asymmetry slowing societal progress and advancement: **business-to-science (B2S) data asymmetries.** As is the case across domains, the private-sector holds massive amounts of data that could provide value for scientific inquiry and research across disciplines. Yet too often that information remains siloed thanks to businesses' concerns regarding competitive advantage and trade secrets, privacy harms or security risks; as well as researchers' lack of recognition of the types of valuable datasets held in the private sector that could support their work, as well as a somewhat dogmatic belief that only data generated in a lab can truly enable new scientific insight. These challenges, as well the relative lack of systematic, repeatable operational and governance models to enable B2S data collaboration lead to persistent transaction costs for the science community related to finding, extracting, formatting, and integrating data to support their analyses,[49] as well as opportunity costs for society at large as achievable, potentially transformative scientific insights continue to go unrealized.
- **North-South and East-West data asymmetries** are another form of data asymmetry and extraction best understood through the lens of data colonialism.[50,51]

ENDNOTES

1 Barnett, Thomas. 2016. "The Zettabyte Era Officially Begins (How Much Is That?)." *Cisco Blogs* (blog). September 9, 2016. https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-that.

2 See, for instance, Bradford Cross, "Big Data is Less About Size, And More About Freedom," *TechCrunch* (2010). http://techcrunch.com/2010/03/16/big-data-freedom/.

3 Mike Loukides, *What is data science?*, (Boston: O'Reilly Media, Inc., 2010). http://radar.oreilly.com/2010/06/what-is-data-science.html.

4 Big Data has also been tagged – more recently – with a more negative connotation, as in parallel to "Big Pharma '' or "Big Oil", as a shorthand for the increased embrace by data brokers of new analytical tools to profile customers and other people.

5 "Community cleverness required," *Nature* 455, no. 1 (2008). https://doi.org/10.1038/455001a

6 Francesco Cappa et al., "Big Data for Creating and Capturing Value in the Digitalized Environment: Unpacking the Effects of Volume, Variety, and Veracity on Firm Performance," *Journal of Product Innovation Management 38*, no. 1 (2020): 49-67. https://doi.org/10.1111/jpim.12545. ; "What is Big Data?", Informatica, Accessed March 28, 2022. https://www.informatica.com/services-and-training/glossary-of-terms/big-data-definition.html. ; Victoria Rubin and Tatiana Lukoianova, "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?", *Advances In Classification Research Online 24*, no. 1 (2013): 4-15. http://doi.org/10.7152/acro.v24i1.14671.

7 See "Technology: The data deluge," *The Economist* (2010). https://www.economist.com/leaders/2010/02/25/the-data-deluge?story_id=15579717.

8 See for instance: Anil Ananthaswamy, "I, algorithm: A new dawn for artificial intelligence," *NewScientist*, January 26, 2011. https://www.newscientist.com/article/mg20927971-200-i-algorithm-a-new-dawn-for-artificial-intelligence/?ignored=irrelevant ; Anthony Bagnall et al., "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery 31* (2017): 606–660. https://doi.org/10.1007/s10618-016-0483-9.

9 See Technology: The data deluge," *The Economist* (2010). https://www.economist.com/leaders/2010/02/25/the-data-deluge?story_id=15579717.

10 See Danah Boyd and Kate Crawford, "Six Provocations for Big Data," A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society (2011): 3. http://dx.doi.org/10.2139/ssrn.1926431.

11 Dale Neef, *Digital Exhaust: What Everyone Should Know About Big Data, Digitization and Digitally Driven Innovation*, New Jersey: Pearson Education (2014); Gerard George et al., "Big Data and Management," Academy of Management Journal 57, no. 2 (2014). https://doi.org/10.5465/amj.2014.4002.

12 José van Dijck, "Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology," *Surveillance & Society 12*, no. 2 (2014): 197-208. https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/datafication/datafic.; W. Weerkamp and M. de Rijke, "Activity Prediction: A Twitter-based Exploration," SIGIR 2012 *Workshop on Time-aware Information Access* (2012). https://hdl.handle.net/11245/1.381106.

13 Ulises A. Mejias and Nick Couldry, "Datafication," *Internet Policy Review 8*, no. 4 (2019). https://policyreview.info/concepts/datafication.

14 Nick Couldry and Ulises A. Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media 20*, no. 4 (2018): 336-349. https://doi.org/10.1177/1527476418796632.

15 Marina Micheli et al., "Emerging models of data governance in the age of datafication," *Big Data & Society 5, no. 2* (2020). https://doi.org/10.1177/2053951720948087.

16 Jens-Erik Mai, "Big data privacy: The datafication of personal information," *The Information Society* 32, no. 3 (2016): 192-199. https://doi.org/10.1080/01972243.2016.1153010.

17 José van Dijck, "Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology," *Surveillance & Society 12*, no. 2 (2014): 197-208. https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/datafication/datafic.

18 Jathan Sadowski, "When data is capital: Datafication, accumulation, and extraction," *Big Data & Society 6*, no. 1 (2019). https://doi.org/10.1177/2053951718820549.

19 Nick Couldry and Ulises A. Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media 20*, no. 4 (2018): 336-349. https://doi.org/10.1177/1527476418796632.

20 Ulises A. Mejias and Nick Couldry, "Datafication," *Internet Policy Review 8*, no. 4 (2019). https://policyreview.info/concepts/datafication.; Prabhakar Krishnamurthy, "Towards Data Science," *Towards Data Science*, September 12, 2019, https://towardsdatascience.com/survey-d4f168791e57; Nicol Turner Lee and Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," *Brookings* (Brookings, May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/; Joy Lu et al., "Good Explanation for Algorithmic Transparency" (November 11, 2019). http://dx.doi.org/10.2139/ssrn.3503603.

21 Mikkel Flyverbom, et al., "The Governance of Digital Technology, Big Data, and the Internet: New Roles and Responsibilities for Business," *Business & Society 58*, no. 1 (2019): 3-19. https://doi.org/10.1177/0007650317727540.

22 Leigh Dodds, "What is data asymmetry," *Lost Boy* (2017). https://blog.ldodds.com/2017/03/24/what-is-data-asymmetry/#:~:text=The%20term%20data%20asymmetry%20refers,more%20value%20than%20a%20contributor; OECD, *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies* (Paris: OECD Publishing, 2019), https://doi.org/10.1787/276aaca8-en; *ODB Global Report Third Edition*, (Geneva: World Wide Web Foundation, 2015), https://opendatabarometer.org/3rdedition/report/.

23 Ulises A. Mejias and Nick Couldry, "Datafication," *Internet Policy Review 8*, no. 4 (2019). https://policyreview.info/concepts/datafication.

24 Piketty, Thomas. 2022. A Brief History of Equality: Cambridge, MA: Belknap Press.

25 Stefaan G. Verhulst et al., *The Emergence of a Third Wave of Open Data: How To Accelerate the Re-Use of Data for Public Interest Purposes While Ensuring Data Rights and Community Flourishing*, "Brooklyn: The Governance Lab, 2020). http://dx.doi.org/10.2139/ssrn.3937638.; Andrew Young et al., "The Third Wave of Open Data: Connecting the Past, Present, and Future of Re-Using Data to Advance Societal Goals," *The GovLab Blog*, The GovLab (2020). https://blog.thegovlab.org/post/the-third-wave-of-open-data.

26 Verhulst, Stefaan, Andrew Zahuranec, Andrew Young, and Kateryna Gazaryan. 2021. "The Third Wave of Open Data Toolkit: Operational Guidance on Capturing Institutional and Societal Value of Data Re-Use." Brooklyn, New York: Open Data Policy Lab. http://files.thegovlab.org/The-Third-Wave-of-Open-Data-Toolkit.pdf.

27 Stefaan G. Verhulst et al., Leveraging Private Date for Public Good: A Descriptive Analysis and Typology of Existing Practices, (Brooklyn: The Governance Lab, 2021). https://datacollaboratives.org/static/files/existing-practices-report.pdf.

28 Yves-Alexandre de Montjoye, Jake Kendall, and Cameron F. Kerry. "Enabling Humanitarian Use of Mobile Phone Data." *Issues in Technology Innovatio*n (November 2014). https://dspace.mit.edu/handle/1721.1/92821

29 Gauvin, Laetitia, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. 2019. "Gender Gaps in Urban Mobility." ArXiv:1906.09092 [Physics], June. http://arxiv.org/abs/1906.09092.

30 Verhulst, Stefaan, and Andrew Young. 2016. "Battling Ebola in Sierra Leone." OPImpact. 2016. http://odimpact.org.

31 Gauvin, Laetitia, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. 2019. "Gender Gaps in Urban Mobility." *ArXiv:1906.09092* [Physics], June. http://arxiv.org/abs/1906.09092.

32 "Data Collaboratives Home Page." 2018. Data Collaboratives. 2018. http://datacollaboratives.org/explorer.html.

33 Verhulst, Stefaan G. 2015. "Data Collaboratives: Exchanging Data to Improve People's Lives." *Medium* (blog). April 25, 2015. https://sverhulst.medium.com/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a.

34 Michelle Winowatan, "The Emergence of Data Collaboratives…in Numbers," The GovLab (blog), accessed April 1, 2022. https://blog.thegovlab.org/post/the-emergence-of-data-collaboratives-in-numbers.

35 A market-led compensation model may address these concerns best. Such a model should consider all associated costs, such as infrastructure setup and maintenance (e.g., API development), technical and administrative expenses, and more. It should also include provisions for reinvestments and innovation, ensuring that organizations sharing data can enhance their capabilities over time. By adopting such a compensation approach, the sharing of data can be incentivized, leading to improved collaboration and increased accessibility to valuable information.

36 "The 100 Questions", *The GovLab*. https://the100questions.org/.

37 A full list of areas and questions can be found here: https://the100questions.org/about.html.

38 Verhulst, Stefaan. "Unlocking AI's Potential for Good Requires New Roles and Public-Private Partnership Models." *AI+1: Shaping Our Integrated Future* (blog). The Rockefeller Foundation, 2020. Accessed March 22, 2022. https://www.rockefellerfoundation.org/blog/unlocking-ais-potential-for-good-requires-new-roles-public-private-partnership-models/.

39 For more on data stewards, see Stefaan G. Verhulst, "The Three Goals and Five Functions of Data Stewards," *Data Stewards Network,* September 20, 2018. https://medium.com/data-stewards-network/the-three-goals-and-five-functions-of-data-stewards-60242449f378.; and Stefaan G. Verhulst et al., *(Re-)Defining the Roles and Responsibilities of Data Stewards for an Age of Data Collaboration*, (Brooklyn: The Governance Lab, 2020). http://www.thegovlab.org/static/files/publications/wanted-data-stewards.pdf.

40 Zahuranec, Andrew J. 2021. "The '9Rs Framework': Establishing the Business Case for Data Collaboration." Data Stewards Network (blog). November 9, 2021. https://medium.com/data-stewards-network/the-9rs-framework-establishing-the-business-case-for-data-collaboration-26585455ccc0.

41 In the context of data sharing, companies often find themselves in a dilemma. While they may recognize the value of sharing data for the greater good or to foster innovation, they also incur costs in collecting, processing, and maintaining that data. Therefore, they may not necessarily seek to generate direct profit from sharing data but rather expect fair compensation to cover the expenses associated with making data available. Data collection and management require significant investments in infrastructure, data governance, security measures, and skilled personnel. In addition, there are legal and regulatory compliance costs to consider. Companies may view data as a valuable asset that deserves compensation, particularly when it can be used to fuel AI advancements or drive insights that create commercial value. Fair compensation models can help address the costs incurred by data providers and incentivize them to share data more willingly. Such models could involve licensing agreements, data marketplaces, or collaborative partnerships where the value derived from data is shared among stakeholders.

42 Aitken, Mhairi, Ehsan Toreini, Peter Carmichael, Kovila Coopamootoo, Karen Elliott, and Aad van Moorsel. 2020. "Establishing a Social Licence for Financial Technology: Reflections on the Role of the Private Sector in Pursuing Ethical Data Practices." *Big Data & Society* 7 (1): 2053951720908892. https://doi.org/10.1177/2053951720908892.
Choo, Mabel, and Mark Findlay. 2021. "Data Reuse and Its Impacts on Digital Labour Platforms." *SSRN Scholarly Paper.* Rochester, NY. https://doi.org/10.2139/ssrn.3957004.
Muller, Sam H. A., Shona Kalkman, Ghislaine J. M. W. van Thiel, Menno Mostert, and Johannes J. M. van Delden. 2021. "The Social Licence for Data-Intensive Health Research: Towards Co-Creation, Public Value and Trust." *BMC Medical Ethics 22* (1): 110. https://doi.org/10.1186/s12910-021-00677-5.

43 Belenguer, Lorenzo. 2021. "Citizens' Assemblies for a More Democratically Engaged and Informed Society on AI & Data Privacy." *Escapadas Ideas Mag* (blog). January 23, 2021. https://medium.com/escapadasuk/citizens-assemblies-for-a-more-democratically-engaged-and-informed-society-on-ai-data-privacy-4d22e1c5b585.

44 Shaw, James A., Nayha Sethi, and Christine K. Cassel. 2020. "Social License for the Use of Big Data in the COVID-19 Era." *Npj Digital Medicine 3* (1): 1–3. https://doi.org/10.1038/s41746-020-00342-y.

45 Choo, Mabel, and Mark Findlay. 2021. "Data Reuse and Its Impacts on Digital Labour Platforms." *SSRN Scholarly Paper.* Rochester, NY. https://doi.org/10.2139/ssrn.3957004.

46 Aitken, Mhairi, Ehsan Toreini, Peter Carmichael, Kovila Coopamootoo, Karen Elliott, and Aad van Moorsel. 2020. "Establishing a Social Licence for Financial Technology: Reflections on the Role of the Private Sector in Pursuing Ethical Data Practices." *Big Data & Society* 7 (1): 2053951720908892. https://doi.org/10.1177/2053951720908892.

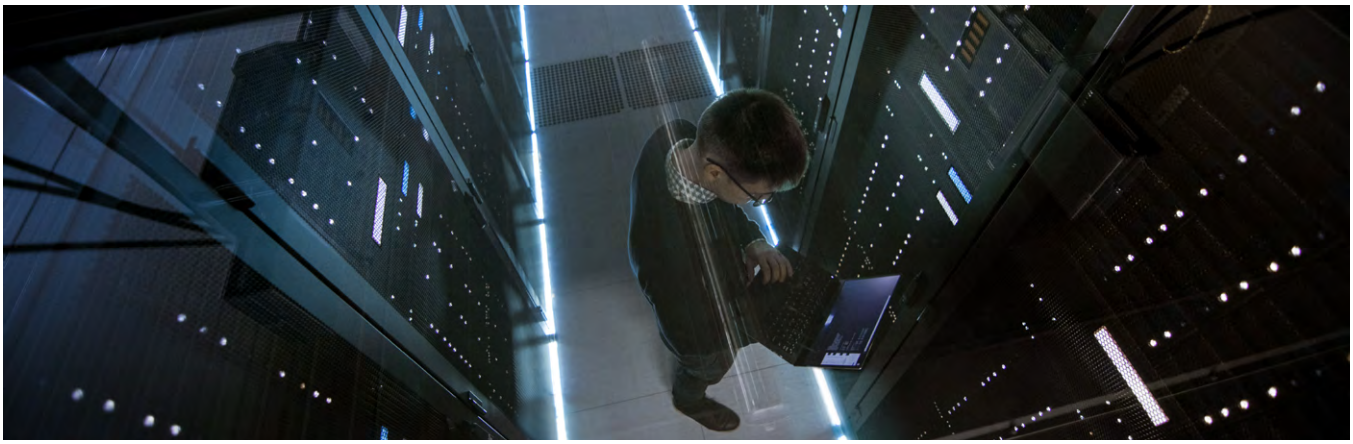
47 For a list of some projects undertaken by the GovLab toward achieving these ends, see here: https://apolitical.co/solution-articles/en/to-turn-the-open-data-revolution-from-idea-to-reality-we-need-more-evidence.

48 "Commission appoints Expert Group on Business-to-Government Data Sharing," Shaping Europe's digital future, European Commission, November 22, 2018. https://digital-strategy.ec.europa.eu/en/news/commission-appoints-expert-group-business-government-data-sharing.

49 Barend Mons, "Invest 5% of research funds in ensuring data are reusable," *Nature 578, no. 491* (2020). https://doi.org/10.1038/d41586-020-00505-7.

50 Nick Couldry and Ulises A. Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media 20, no. 4* (2018): 336-349. https://doi.org/10.1177/1527476418796632.

51 Stefania Milan and Emiliano Treré, "Big Data from the South(s): Beyond Data Universalism," *Television & New Media 20, no. 4* (2019): 319-335. https://doi.org/10.1177/1527476419837739.



IE UNIVERSITY

CENTER FOR THE GOVERNANCE OF CHANGE

AUTHOR:

Dr. Stefaan G Verhulst
IE University

RECOMMENDED CITATION:

Verhulst, S., *"Data Collaboratives: Enabling a Healthy Data Economy through Partnerships"*, IE CGC, July 2023

© 2023, CGC Madrid, Spain

Design: epoqstudio.com

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License. To view a copy of the license, visit creativecommons.org/licenses/by-nc-sa/4.0

cgc.ie.edu