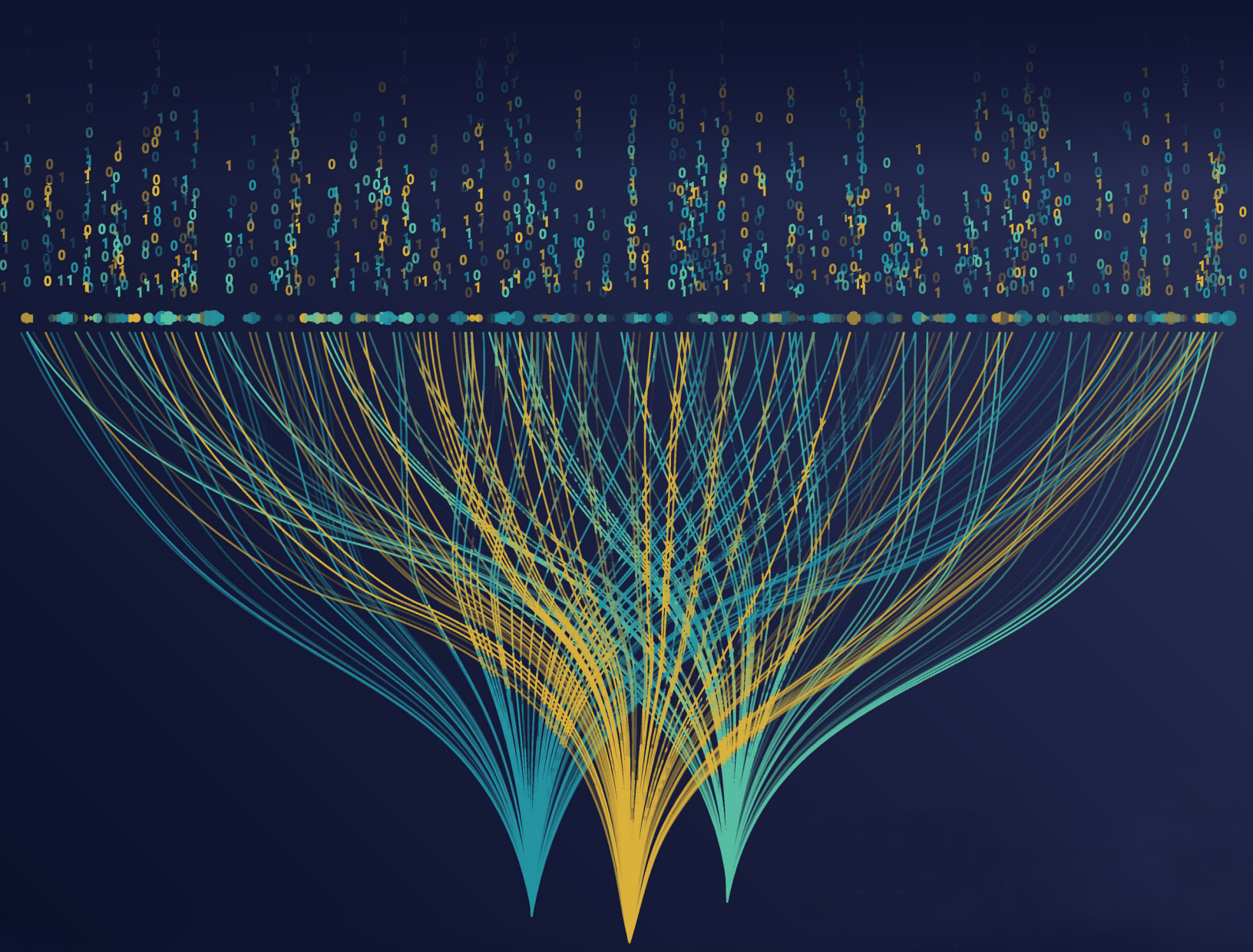


A Data Sharing Discipline

STEVEN WEBER, MATTHEW NAGAMINE, MAX INGRAHAM-RAKATANSKY

UC BERKELEY
SCHOOL OF INFORMATION

AND

UC BERKELEY
CENTER FOR LONG-TERM CYBERSECURITY

A Data Sharing Discipline

STEVEN WEBER, MATTHEW NAGAMINE, MAX INGRAHAM-RAKATANSKY

UC BERKELEY
SCHOOL OF INFORMATION

AND

UC BERKELEY
CENTER FOR LONG-TERM CYBERSECURITY

Contents

Introduction

If data is the most valuable asset an organization possesses, why would the organization consider sharing that data with anyone else?

This isn't a new question. Isaac Newton pushed his 17th-century contemporary John Flamsteed to share his astronomical observations, arguing that because the work had been funded by the Royal Treasury, Flamsteed's data was a public resource that should be shared to accelerate further discovery.[1] Rapid advances in data science have now raised the stakes for developing clear arguments about the precise meaning of data sharing. We need a discipline that defines the reasons why and conditions under which any organization — whether in the public or private sector — should choose, be incentivized, or even be required to share data. A practical set of arguments needs equally to address sharing with whom, under what conditions, and — most importantly — to what purposes.

This discipline will not be solely a calculation about value, profit, property, privacy, security, competitiveness, or public good, but a combination of these and possibly other considerations. We will know we have it right when an explicit data-sharing logic is able to define success, showing that sharing data leads to better outcomes than not doing so on some set of agreed dimensions. Right now, such a discipline around data sharing does not exist, and this paper is an effort to move the debate forward in that direction.

A fully articulated data sharing scheme will address at a minimum three points. What data is in play? What actors (people, firms, government agencies, etc.) are involved and on what terms? And to what ends, goals, and objectives does data sharing contribute? Many contemporary data sharing schemes have partial answers; some have missing, inconsistent, or contradictory ingredients. The aim is not perfection but simply movement toward a better synthesis and fewer disagreements. That requires breaking down the problem in order to reassemble it in new ways.

Strong Priors

Debates about data sharing mobilize strong prior beliefs. One bundle of beliefs starts from the foundational view that data is now replacing intellectual property in some settings as the most potent value-differentiator, as firms that can best collect, manage, and extract insights from data have a distinct competitive advantage. If data is that valuable, why share it with competitors and possible competitors? And why do that right now, in a moment of rapid technology change and explosive growth in tools that promise to extract meaning from massive data sets, a moment where no one can be certain about what actually is present in data and ready to be discovered?

Another bundle of beliefs starts from a different foundational view of data as a collective good or common human heritage that is an essential input to new knowledge and value creation. This perspective mostly rejects the conceptual apparatus of property rights and competitive advantage in data, focusing instead on openness, essential foundations of discovery, pre-competitive collective goods, and related definitions of the "public interest" or "commons."[2] The default starting point here is that data should be shared as widely as possible, and that the burden of proof falls on public- and private-sector organizations that want to enclose data for whatever reason. That burden of proof might be met with arguments about security or privacy, or perhaps even about incentives and efficiency, but the threshold for justifying enclosure should be set at a high level.

These perspectives (and others, which we consider in more detail below) share some starting assumptions. The most important is that data is fast becoming less expensive to collect in many (but not all) cases, while it is still relatively difficult and expensive to organize, maintain, protect, and process for insight. Raw data is cheap, but acquiring quality data sets and generating inferences and insights from data are not. Another shared assumption relates to uncertainty: whether insight will emerge in data sets drawn from particular sources, or at the unexpected intersections between sources, remains highly uncertain.

A final shared assumption cuts across perspectives on what the nature of the data sharing landscape is and should look like. Behaviors and discourse around data sharing now often seem inconsistent. This could simply be the result of confused and incoherent thinking. It could also be intentional and strategic: the baseline for many organizations at the intersection of these multiple uncertainties and diverse, often politicized perspectives might be to talk the talk of openness and sharing, but act to hoard data. That logic could apply just as well to public-sector

and non-profit organizations as to private firms. Organizations may not be competing for profit, but they certainly compete for power and influence in other spheres, where decisions about how to share or not share data can be just as crucial.

Data is not the "new oil."[3] But the complexion of data as a key input to scientific, economic, and other value-creating processes bears a similarity to oil concessions in the early part of the 20th century, in that the relevant players who would invest in extracting value did not really know how to judge in advance what might lie beneath the sands, or what it would take to drill, refine, and get the valuable derivative product to market. Sharing of oil concessions was certainly not a favored strategic response to that situation of uncertainty.

But data sharing is much more common than that, and debates about incentivizing and/or mandating the sharing of data are much more widespread. To better understand this phenomenon, we propose a conceptual framework and a common language for analysis of shared-data initiatives and value creation. This new framework is needed right now, and not just for the sake of academic research. Decision-makers and strategists need to get past ideologies and prior belief sets to confront high-stakes, practical decisions about what data to share, how to share it, with whom, for how long, and — most importantly — to what ends.

Why Now?

The prominence of shared data as a purported solution across many initiatives and proposals begs the question: what exactly is the problem that shared data solves? In mid 2020, there are several different answers to that question, sometimes implicit and sometimes mixed into confusing combinations. These problems include:

- **Lack of precision in calculations about the actual value of particular data and data sets.** From a simple economic perspective, highly imperfect markets for most data sets make efficient and commonly agreed price discovery nearly impossible. We will later consider in more detail how this complicates market exchanges of data that would be value-creating, and leads to calls for sharing as a way to "get around" the market failure problem. A concrete example would be a transportation company that has a sophisticated and granular data set about the demand for mobility within a city, and a city government that would like access to that data for infrastructure planning. An efficient market would enable trade — a straightforward exchange of private data for public-sector resources that

produces value for both parties. The firm would monetize its data, and the government would gain knowledge about where to invest. Because that kind of exchange does not seem to happen as expected, data sharing (in this case, from business to government or B2G) is put forward as a solution to a market failure problem.

- **A complex mix of assumptions and ideologies about public and private goods in relation to data.** This problem is rooted in a first-principles debate about who (if anyone) should possess what property or usage rights in data. When someone walks down the street to buy a drink at their local coffee shop, they create valuable data along the way — but who should be able to claim the right to own or use that data? Language really gets in the way here. It has become commonplace to read statements like "90% of the world's data has been created in the last two years."[4] But this is a peculiar statement because taking a walk to get coffee does not create any more data than it did 20 years ago. The obvious difference is that what we now call data is today *collected* by mobile phones and an array of other sensors along the way. What mix of property rights belongs to the creator and what to the collector? Some ideologies would solve this problem by asserting that "you own your own data," but beyond concerns for privacy, the rationale for that seems strained since the data as such does not exist (at least not in a usable form) without the creation and operation of the sensors and systems that collect it, many of which consumers do not own and did not invest to build and deploy. On the flip side of that same ideology is a contradictory assertion that this data ought to be available to anyone who can use it to create public goods. Some concepts of data sharing would make this debate irrelevant and "solve" the problem by enforcing a shared data model whenever the potential for creating public goods exceeds the potential for creating private harms (in particular, privacy negatives). But finding agreement on how to measure those parameters may shift the problem into another first-principles space that could be even harder to solve.

- **A worldwide drive to increase productivity and entrepreneurship, both of which might now be fueled by easier access to shared data resources.** The search for productivity was in high gear before the COVID-19 crisis, notably but not only in the most advanced and wealthy economies, which have been suffering from tepid economic growth and weak productivity since at least the 2007-2009 global financial crisis. In some formulations, there has been a gradual secular decline in both growth and productivity for decades.[5] As Paul Krugman put it, productivity growth in the long run is not everything, but it is *almost* everything, and despite the U.S. West Coast's self-image as a globally unmatched start-up ecosystem, entrepreneurship rates in the United States have been declining for

quite some time.[6] The COVID-19 crisis is certain to supercharge that drive for productivity as economies around the world desperately seek higher rates of economic growth to manage the debt load that unprecedented emergency rescue measures will leave behind. If the bet is that AI/ML technologies will be one of the most significant engines for growth and productivity in the next decade, there is a strong argument for lowering barriers that restrict experimentation and the widespread entry of new ideas at low cost. It is plausible to imagine (as we will explore later) that shared data might help lower such barriers, as cloud computing did during the last decade. These two domains might even be synergistic: the ability to access computing power and data sets on "utility" terms would provide a good head start for a machine learning scientist who seeks to create a new business. This argument is independent of claims that shared data could be a tool used within competition policies to reduce the power of data-rich incumbents.

These are just a few examples of the kinds of questions or dilemmas to which data sharing is put forward as an answer or solution. It is illuminating to re-categorize these arguments by deep rationales that link them, to illustrate why debates about data sharing sometimes seem stuck in incommensurable paradigms.[7]

One paradigm highlights **moral and ethical stakes**. These arguments often emphasize obligations to the subjects from whom the data came and the potential harms (absolute and relative) they might suffer. They also emphasize the value of knowledge derived from large data sets as a public good in terms of aggregate benefits to society.

Another paradigm highlights **practical scientific upside.** These arguments stress the value of shared data for reproducibility of experiments, and particularly for enhancing the overall efficiency of the collective research enterprise (avoiding the re-invention of wheels, the pursuit of what others have found to be dead-ends, and the like). Some of these arguments focus on the contribution of sharing data to social trust and collaboration among researchers.

A third paradigm highlights **economic and innovation effects.** These are seen as positive values in and of themselves, as well as instrumental underpinnings to the upside potential of the other paradigms. If shared data is understood as a public good, it is probably (like other public goods) vulnerable to under-provision unless the incentives to provide it are rightly configured. It is also true that if the scientific enterprise benefits from risk-taking and innovation, both are clearly dependent on incentives for individuals and organizations to bear the risk. Economic and innovation arguments also address more fully the interesting dimension

of time: what is the logic for data (like intellectual property in patent law) to be private in one time frame, licensed in another, and fully shared in a third?

Even a casual observation of the data-sharing landscape shows a mix of prior beliefs and arguments that are often ambiguous on their own, and even more complicated and incommensurable in combination. No wonder the trade-offs are tough to define and even tougher to adjudicate.

Where you Stand Depends (Partly) on Where you Sit

Because discussions about data sharing can be interpreted through many different frames, stakeholders may have a reflexive instinct about what is at stake and what should be done. These baselines should be treated as hypotheses only, but they reflect a simple "where you stand" logic that is a reasonable place to start.

Investors — those who have put resources into collection and preparation of data sets, anticipating economic returns — would start from the presumption of "hoarding," as in the oil concessions argument. That there might be potential for even greater value creation in shared data will not dislodge that instinct, until and unless there are clear mechanisms for ensuring that a high proportion of the value lands with the investors who placed bets to start.

Researchers and scientists who use data as an input to their work (including but not limited to machine learning) would start from the presumption that data shared and pooled in a standardized manner best serves their interests, because it reduces the cost of inputs to the research enterprise. But this works only if their version of return on investment (ROI) — receiving credit for data collection and preparation, and for findings and products that make use of shared data — can be protected and rewarded.[8]

Privacy advocates would likely start from a more skeptical place. All things equal, the sharing of data will increase the vulnerability of that data to mishandling and potential privacy harms. As more parties gain access and data with different privacy permissions mix, the likelihood of privacy compromise will rise. So privacy advocates will prefer to minimize data collection

and preservation, lock down data sets, keep them separate from each other (physically and logically), add controls, and reduce single points of failure. They will demand higher levels of proven value creation in data-sharing schemes to justify the increased privacy risks.

Government agencies, taken as a group, would tend to favor data-sharing schemes that provide easy access to data when they want it, to enable government action and service provision. Governments, of course, employ scientists, privacy advocates, and other professionals, and governments vary in the weight they place on different values around data. But when data that would benefit government action is collected by the private sector (and particularly by several private-sector players, not just one or two), governments' appetite for B2G (business-to-government) sharing will rise. At this level of abstraction, it does not matter if the intended action is pandemic response and disease surveillance, criminal investigation and prosecution, granular economic activity assessments, or the identification and persecution of political opponents. At the same time, governments vary in how they initiate (or do not initiate) sharing of their data assets, to other governments or to businesses. (Open data initiatives will be discussed in further detail below.)

Individuals and organizations (including NGOs) that focus on the provision of public goods often view data sets first and foremost as a source of inputs to their mission. This is parallel to for-profit firms that would naturally prefer open-access, low-cost inputs to their production processes, but different in terms of preferences on the output side. Focusing on the potential for shared data to energize public goods provision has led to the proposal of umbrella concepts like FAIR ("findable, accessible, interoperable, reusable"), a seemingly common-sense structure similar to the RAND ("reasonable and non-discriminatory") licensing framework used in relation to patents.[9] But the FAIR concept does not apply everywhere, nor does it by itself answer the question of what data should be shared, when, and with whom.

End-users, citizens, and customers who access products and services that depend on shared data are stakeholders as well, but their starting beliefs are harder to establish by hypothesis. The typical end-user of a commercial data product (for example, transportation maps and guides) or a government data product (for example, a detailed employment report from a statistical agency like the Department of Commerce) will not have a clear line of sight into how data sharing enabled the creation of that product; nor will they be able to assess a counterfactual scenario of what the product would look like without a data-sharing scheme behind it. So the starting point for this group of stakeholders will likely depend on narratives that others create around shared data. This comes on top of the different starting positions

they inhabit with respect to how inferences and products that rely on shared data would impact them. For example, a property owner in a wildfire zone might be more willing to tolerate the sharing of private data among neighbors than would a property owner in a less fire-prone area.

This exercise of establishing hypotheses about priors and baselines could be extended widely to a range of stakeholder groups, such as property rights fundamentalists, small and large firms, and others. The diversity of stakeholder priors creates another layer of incommensurable arguments with largely unmeasured parameters. How can a government agency, a privacy advocate, an investor, and an end-user argue productively and come to a negotiated agreement about where, when, and how it makes sense to share data?

In most cases, they cannot. To make progress toward shared understanding about shared data that might improve those arguments, we need better starting definitions of what shared data is and what it is not, and that is the goal of the next section.

Definitions Matter

"Sharing" and "data" are commonly used words that have highly variable meanings in different settings. In casual parlance, "sharing" can mean to *use* something jointly, or to *give* a part or whole of something to others. Examples in the physical world include sharing a house (through usage or ownership), ride-sharing through a service like Lyft (which is more like purchasing a unit of transportation from point A to point B), and sharing access to public goods like parks and highways, where the meaning is embedded in the "non-excludable" aspect of public goods. In the digital world, sharing can also mean *conveying* information or passing it along to others, such as a Facebook status update, sharing media through peer-to-peer file sharing, or *collaborating* on documents using a tool like Google Docs.

For the purposes of this paper, "data" can be quantitative or qualitative, raw or highly processed. It may be derived from any number of sources, ranging from IoT devices in a home to online transactions or biosensors, to a government census, to the transcripts of a focus group discussion. Datasets are collections of data, whether a set of economic and cultural indicators, geospatial data, or genomics data from DNA sequencing.

To explain data-sharing practices with common-sense definitions like these is impossible, as there are too many degrees of freedom within each term and too many possible permutations. We need to limit the scope in some fashion. No definition is perfect, and the usefulness of any definition will evolve with changing practices, but for this paper, we bound the concept of "shared data" with five constraints.

For our purposes, the definition of shared data *is not*:

- the sharing of data between an individual and a company, as in colloquial expressions like "I shared my data with Google." Such transfer of data is more of a conventional exchange, a simple trade of data from an individual to a company in return for a service. If the company monetizes that data through third-party advertising, that is a business model choice that sits alongside, but does not fundamentally alter the fact of exchange.

- the same as "open data." To grant open data rights generally means giving data away to anyone, for any purpose. Examples include the World Bank's open data sets or municipal efforts to open real-time transit information to app developers. The open-closed dimension that describes a data set is independent from sharing. While shared data can be made open in this respect, it can also be made closed, thus it makes sense to distinguish sharing from openness.

- a public domain "data donation." The act of donating data to the public domain, particularly if the data is unstructured and the donator eschews responsibility for the quality of the data or managing it going forward, does not fall into our definition of sharing.

- a "transparency initiative," which is typically the project of a third party (i.e. not the data creator or presumptive owner) who wants to render someone else's data visible for some purpose. An example would be making financial contributions to political campaigns visible to discourage corruption, i.e. "disinfection by sunlight." Transparency may create value and improve accountability, but through a different mechanism that generally involves holding people, firms, and governments to account. If a transparency initiative involves unilateral appropriation of data, that is clearly not an exchange. If the appropriation is illegal, the best word to describe it is simply "theft."

These constraints bound the problem by saying what shared data *is not*. We use four characteristics to define that shared data *is*:

- the product of collective action between two or more players. Sharing is not simply a one-to-one exchange, but a joint and intentional act of a meaningful number of players ("meaningful" in the context of their market). It also implies that shared data initiatives will often be subject to some manner of collective action problems.

- the bringing together of data sets that each player owns or has access to, but others do not, to create more than the sum of parts. Sharing creates value.

- based upon some agreed framework that specifies operational format and structure, including how the shared data will be standardized, curated, stored, updated, kept clean, etc.

- an agreement about the most important rights and responsibilities of the sharing arrangement.[10] Who contributes what? Who can take or use the shared data? What can be done with it and to whom does the consequent value belong? The next section draws out a more granular typology along these lines.

These definitions are not absolute, ideological, or theological in nature. Rather, they provide a pragmatic means of bounding the problem for the sake of analysis. The terms of the definition might evolve over time and the boundaries might soften in practice. But adopting this more precise definition of data sharing should make it easier to understand the stakes, and help bring clarity to relevant policy and strategy debates.

A Preliminary Operating Typology

Data sharing needs a good map to assess the landscape, but how the map should be organized is not self-evident. Mapping by economic sector would make sense if we assume that the interesting dimensions of variance in data sharing practices are associated with distinct economic sectors (that is, if we expect different practices by energy companies, health care firms, or government). But that would ignore what is most important about data and data flows: they cut across sectors and do not naturally sort along industry lines.

A second approach to mapping might start with the hunch that the most interesting variance lies in "types" of data — PII vs. industrial IoT data, for example. But that would require a strong hypothesis about how to segment variations in types of data. While privacy sensitivity is an obvious approach, it is unlikely to be the most important. If it were, we would observe greater data sharing in privacy-insensitive domains than in areas where privacy concerns are substantial, and that is not the case. We would also expect that privacy concerns around data sharing would be much less prevalent than they are, as organizations' practices would have already sorted around these concerns.

The approach here starts instead with a simple observation about three dimensions that vary among data-sharing schemes in practice. The rationale is an assumption that organizations make decisions about data sharing strategically and under uncertainty. And so the choices they make reveal fundamental dimensions of variance in the landscape that shape those choices. It's important to keep in mind that a typology is structured description, not theory. The typology does not explain why a particular data-sharing scheme shows up in a specific spot in the landscape (any more than a political map explains why the borders between countries are where they are). The typology rather serves as a starting point for organizing what needs to be explained, and sometimes points to hypotheses about where those explanations might be found.

Three strategic dimensions make up the typology:

1. *Where* in the value chain are organizations sharing data?

This first dimension is likely the most intuitive. Data gets created continuously along a value chain. Most organizations segment their value chain by how it relates to competition in the marketplace, and this dimension follows that segmentation.

Note that the boundaries are sometimes fuzzy. Consider, for example, the distinction between "basic research" and "pre-competitive applied research" in today's pharma sector, enterprise software, or public-sector service provision. The lines will often shift over time, as technology and business models experiment and evolve. Organizations sometimes try to move the

line intentionally to gain advantage, for example by aiming to commoditize their inputs and concentrate competition in parts of the value chain where they are strongest. Even with those complications, the value chain dimension suggests an intuitively attractive starting hypothesis that data sharing is easiest to organize and most likely toward the left side of the axis (in basic research), and becomes progressively harder and less likely as it moves to the right.

The mini-cases in this paper demonstrate this effect, but also show that other considerations can strongly modify the expected pattern. For example, post-competitive data sharing in the pharmaceutical industry functions as a *de facto* stage-4 clinical trial, when a drug is widely used and large-scale data emerges about efficacy, side effects, and drug interactions. It is clear that data sharing in that segment is to some degree mandated by regulators as a public good.

The cases also show unrealized potential for value creation in the pre-competitive segment. For many machine learning research projects, the limited availability of large, high-quality data sets acts as a significant barrier to entry. Researchers with fewer resources and small start-up firms are held back; large incumbents with a head start in collecting data are advantaged. This is one of the most prominent arguments in competition policy debates about natural monopoly associated with certain machine learning technologies.[11]

Shared data in the pre-competitive phase could become this decade's conceptual equivalent of Amazon Web Services' "elastic computing," which vastly reduced barriers to entry in the 2010s. Shared data is certainly more complicated than elastic computing resources: data isn't nearly as general and all-purpose as cloud computing; it is not straightforward to turn access on and off, for example. But these are not immovable constraints, rather they are business model problems to be solved by imagination and negotiation. The potential for mobilization of "elastic data" in the pre-competitive phase shows how data-sharing arrangements could contribute to a more vibrant competitive and innovative ecosystem in both commercial and non-commercial machine learning work.

2. *What* can organizations do with shared data?

The second dimension in the typology of data schemes relates to what players are empowered to do with shared data. Shared data has permissions attached to it, as would any dataset, including one owned and used by a single organization. How a dataset is *created* does not determine the ways in which it can be *used;* the two dimensions are distinct. The players who share data can agree in principle to whatever rights and restrictions they wish.

This is most easily seen from a property rights perspective, where the analogy to distinctions among open-source software licenses is revealing (as shown in chart 2). BSD-style licenses have the fewest restrictions: you can take BSD code and do just about anything you want with it, including package, sell, and build on top of it with proprietary code that remains proprietary. The analogy is simple: a firm could use data from a shared dataset, build machine learning products from it, sell those products in a fully proprietary setting, and incur no obligations back to the shared dataset or the players that generated it.[12]

But those kinds of obligations could just as easily be mandated in a different property rights regime. The GPL-style license carries with it a crucial restriction: code that uses or builds on GPL code must itself be released under a GPL license. You can build, package, and sell whatever GPL code you like, but code you create in the course of doing that flows back to the open source community under GPL.[13] An analog is a firm that uses shared data and, in the course of doing so, creates new data. Under a GPL-style model, the new data would have to be shared under the same terms as the data that was used to help create it.

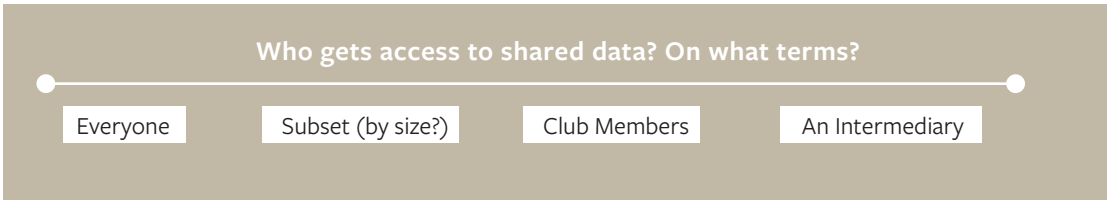
As with GPL developers, a group of firms might agree to such terms for a number of reasons: ideology (a belief in sharing for sharing's sake, or reciprocity) or an assessment of positive externalities resulting from shared data creating aggregate value (that is, the value of shared data grows faster as more data is shared). It might also be a defensive move to reduce the possibility that a non-contributing player could realize a gain from using shared data without providing a benefit to those firms that contributed to the dataset.

The property rights perspective is not the only one that could structure restrictions on what organizations can do with shared data. Privacy represents another model, as does geography, as shared data could be licensed for use within a set of political borders, and not beyond. Of course, none of this says that enforcing these kinds of restrictions would be easy or rational. Indeed, variants of "what can you do with the shared data" are limited only by the imagination of those who construct the rules. While open-source regimes are possible, it is equally easy to imagine rules that tightly control what parties to a shared database can and cannot do, similar

to a contract that tries to anticipate and regulate as many contingencies as the parties can foresee.

There should be no preconceptions about where on this continuum the greatest value — social or individual — will be found. Ideologies sometimes default to one or another endpoint, but that's just ideology, not evidence. It should also be clear that this second dimension — what you can do with the shared data — is independent from *where* in the value chain data is shared. A scheme could just as easily impose GPL-style restrictions on any part of the value chain as it could require precise contracts for exchange. Even with only these two dimensions in place, there would be a lot of variation to be explored. But there is a third dimension that must be considered.

3. *Who* gets access to shared data — and on what terms?



The third dimension of sharing delineates who has access to the shared data, and on what terms. The variance here is similarly broad and experimental. Organizations might create a shared dataset for their own use, then offer it up to the world at large; or they might choose to limit access in any number of ways. If big players in a sector contribute the shared data, they might limit access only to other big players as a way of sustaining oligopoly (small players, shut out from the shared data, would presumably then suffer disadvantage). Another model, based on reciprocity, might offer access to a shared dataset only to those who contribute something valuable to it.[14]

This dimension highlights the role of a shared database intermediary, which might not own the shared data, but could control access rights and regulate who is able to read shared data, as well as write to it. Such institutional arrangements may be most likely to arise where privacy concerns loom large — for example, in shared health datasets.

The access dimension has a price component layered into it, but price and access are not always correlated. Everyone could have access to a shared dataset, but not necessarily at zero price (consider the FRAND principle in IP regimes).[15] Access could be limited to an oligopoly

with the price being zero, or high enough to reinforce the oligopoly rights, or somewhere in between. And price could overlap with limiting access to those who contribute, i.e. the price of access is data reciprocity. The range of variance for who has access on what terms is limited only by the imagination of the players who construct the rules.

• • •

No typology is perfect or enduring, particularly when aiming to categorize a fast-changing set of practices. The point of this typology is simply to begin mapping the landscape in which data sharing practices are likely to differ. A future step would be to plot an appropriate sample of existing data-sharing practices on a three-dimensional grid defined by this typology, to examine where the high-frequency or high-impact clusters are found, and to isolate why those clusters land where they do. Ultimately, the goal would be to experiment with testing the consequences in different contexts — to validate the causal arguments, but more importantly to assess whether new value can be created by sharing data in different and non-intuitive ways. This emerging area has significant potential for innovation.

Models of Value Creation in Shared Data

The primary objective of data sharing should be value creation. That is not to assert that other objectives, such as enhancing competition or protecting privacy, are not important. Rather, it is an assertion that reflects the vast promise of data science to create useful knowledge, as well as the global need to boost economic productivity. Value creation can incorporate important elements of time horizons and sustainability as it serves those goals. After all, a short-term value creation scheme that reduces competition will most likely destroy value over the long term; a scheme that violates privacy considerations will do the same. Ideologies asserting that data ought be shared for normative reasons do not meet the threshold.[16] There are several concrete arguments about value creation — with logics that are distinguishable in the abstract, even if they overlap in practice — that underpin current experiments in shared data.

1. Organizations share data in order to *provide collective goods*. When multiple players recognize that pieces of the puzzle they wish to solve are "owned" separately, a reasonable

response is to bring those pieces together in a shared dataset. During a humanitarian disaster, for example, sharing basic information on meetings, contacts, and resources from the many relief agencies on the ground is essential to the collective good of effective response. Trying to arrange specific trades of particular data needed at a given moment is inefficient relative to a broader data-sharing scheme. As with other collective goods, the logic of under-provision due to externalities is an important reminder that the promise of a collective good by itself is not enough to ensure that organizations will engage in collective action.[17]

2. Organizations share data to *prevent collective harms*. Abstractly, the rationale is equivalent to that of attaining collective goods. In practice, the prevention of collective harms often seems easier to achieve. This may result from human predilections for loss aversion (preventing a harm feels more urgent than achieving the promise of a good); or from how decision-making systems inside organizations treat downside versus upside risk. For example, industry and law enforcement may share data about retail thefts to detect crime patterns and reduce losses. The cybersecurity sector is similarly motivated, but that example also illustrates that the prevention of collective harms through data sharing is constrained by practical considerations like legal liability, competitive pressures, trust deficits, and the technical demands of ensuring the quality of shared data.[18]

3. Organizations share data *to achieve adequate dataset scale or to search for nascent insights*. Many data projects do not scale linearly. In some cases, a dataset will not yield much value until it crosses a certain threshold, at which point a significant amount of value can be realized. To get to that threshold might require investment and effort that no single player can efficiently expend. This is a common issue in pharmaceuticals, where statistical significance in clinical trials requires large experimental datasets that are extremely expensive to produce. Clinical trials also generate lots of additional data that could prove useful in the future or in other settings but have no direct value to the current experiment. This is a common pattern: mining firms create lots of geological data that is not immediately relevant to the extraction of precious metals, and IoT traffic systems generate data that is not useful for timing traffic light cycles, but might be useful for other purposes. The rationale for nascent-value data sharing might be as simple as "I'm not going to do anything with this data, maybe others can." Of course, this rationale by itself will not necessarily motivate the additional effort it takes to organize and maintain a data-sharing scheme; it would be simpler for the organization to simply donate the data, dump it into the public domain, or sell or exchange it in a market setting. Developing a data-sharing scheme under this rationale requires considerable additional effort and thus needs to be justified and explained.

4. Organizations share data when third parties on whom they depend *exert pressure*. Funders of non-profit organizations, for example, may demand that their grantees share data for the purpose of multiplying the efficacy of their grants. Large investors could demand the same of firms in their portfolios. Governments could do the same for their agencies, contractors, or grantees. In the humanitarian relief sector, data sharing may be a response to UN agencies, donors, and recipients, each exerting pressure on independent relief agencies to better coordinate and upgrade their collective capacity and performance. The concept of "data philanthropy" largely fits here, as the UN and the NGO community explore how privately-owned data (such as mobile phone records indicating refugee displacement, or social media indicators of rising food prices) could be shared for aggregate use by humanitarian relief organizations.

5. Organizations share data as a form of *strategic behavior in competitive settings*. Sharing is not equivalent to altruism and is not always associated with friendly intentions toward others. Sharing data is a *means to shape markets*, and can be a powerful tactic to that end. Organizations can share data to undermine other players' competitive advantage, by commoditizing parts of the value chain where other players are dominant. (The open source analogy is relevant here, as well. Making source code free commoditizes parts of the software value chain and shifts the locus of competition elsewhere. The spread of Linux distributions in web servers and data centers, as well as Android's strategy vis-a-vis iOS, are examples.) Shared datasets can empower one side in a negotiation by helping the weaker parties address an asymmetry of power — for example, if employees share information about salaries. A group of small players might share data in order to gain leverage against a very large player in their market, as a less demanding alternative to consolidating or fully merging for the same purpose. For example, the online advertising market might be reshaped by smaller players who share customer profile data in an effort to gain leverage against the large incumbents. This is just a partial list of market-shaping strategies that shared data can empower; it is certain that others will emerge as data economies and ecosystems evolve.

These logics are neither mutually exclusive nor comprehensively exhaustive, but they capture important differences that shape observed practice. Future work could assess the frequency and intensity of each of these logics as drivers of data-sharing. Is one logic more powerful in changing practices, or observed more frequently than others? Are there trends in which driving logics are more prominent over time, or in different sectors or geographies? Are some logics stronger and more sustainable over time; can they support evolving data-sharing schemes or

do they run up against diminishing cost-benefit returns at some point? These are the kinds of questions that researchers and practitioners will encounter going forward.

Practical Considerations: How To Make It Work

Sharing data is a business logic decision. Even the most compelling case for value-creation in shared data is not by itself enough to insure that players in control of separate data sets will shift to a different set of practices to realize the potential value in sharing. The full range of practical roadblocks will become visible over time as experiments in shared data continue, but three core considerations are already visible, along with some insights about how they can be overcome.

The most obvious consideration is how to manage *the distribution of costs and benefits*. An industry that does not share data (equilibrium X) produces a certain amount of value that is distributed unequally among the players. Assume that moving toward shared data practices (equilibrium Y) would create new aggregate value. Equilibrium Y will have a different distribution of value—and, in for-profit sectors, a different landscape of profits. But increasing the size of the pie as a whole does not necessarily increase the size of the slices for all players, and is almost certain not to do so proportionately. Some players will probably see their slices reduced as advantages they previously enjoyed are removed.[19] If the relative losers in the new distribution have the capacity to impede or block the move to equilibrium Y, they may well try to do so. At a minimum, they should be expected to use the threat of obstruction as leverage to bargain for a re-balancing of benefits in their favor.

This is just conventional business strategy, normal blocking and tackling for relative advantage. But the stakes in data at this moment may exacerbate the distributional issues, simply because of the uncertainty about how much value shared data will generate, and for whom. Uncertainty in this respect is probably the enemy of data sharing, and can have pernicious effects in surprising places. The Global Pulse Initiative, for example, has since 2011 used its home at the United Nations as a platform to advocate for what it calls "data philanthropy" in a variety of humanitarian and global commons domains, with "innovation labs" and a variety of tools, but its efforts have run up against almost every roadblock outlined earlier.

A second strategy consideration is also familiar: the energy of activation. There is a time horizon to investing in shared data, and players may not be in a position to make substantial investments ahead of revenue or value that can be released. The humanitarian sector has poignant illustrations of this roadblock. We might want shared datasets and associated structures in place before a disaster hits, but it is a struggle to put up money and effort in anticipation of a need that will come at an undefined future time. Public health authorities around the world are in 2020 living that agonizing reality with regard to the COVID-19 pandemic.

Solutions to this challenge are also familiar. One solution is concentration and size — for example, when one player (or a very small number of tightly coordinated players) is big enough to manage the upfront investment, regardless of what smaller players do.[20] Another solution is external subsidy: a government, a philanthropy, or some other outside party can put forward the energy of activation and drive sharing forward (though that outside party may later want to take a piece of the shared value). This approach is central to the European Commission's data strategy proposal of February 2020.[21]

The third consideration is the role of intermediaries, which extends the energy of activation to other enabling functions that make data sharing schemes easier and more valuable. What kinds of enabling functions would an intermediary perform with the greatest impact? Examples from pharma, energy, advertising, and humanitarian data schemes suggest a few possibilities.

- *Quality control and assurance:* Data, like other inputs, varies in quality. In a shared dataset, it is possible for corrupt, incomplete, or inaccurate data to do more damage, rather than be damped out.[22] An intermediary can serve as a neutral source of quality control and assurance that serves the sharing community as a whole.

- *Data engineering, warehousing, and management:* Managing data at a certain scale becomes a difference not just of size, but of kind. An intermediary can step in to engineer and maintain the new system that will be needed. How will data be configured and organized? What permissions attach to particular actions (such as writing to and reading from the shared data)? What about authentication, authorization, and other aspects of security engineering? And what is the time dependence of the data? How urgently must it be updated and how quickly does it obsolesce?

- *Provision of shared tools:* Shared data often benefits from or requires expensive and technologically sophisticated tools. If building those tools is difficult or expensive, their availabil-

ity upfront may be another challenge to activation. Subsidies can solve for that, but so can an intermediary with a direct interest in taking some of the value released.

Every strategy argument about standards also applies to shared data. Decisions about standard-setting are high stakes, and are almost never a neutral engineering decision, because standards have distributional consequences that benefit some players more than others. Standards also represent a bet (sometimes a very early bet) on where value lies in data and how best to access that value. A deficient set of standards could cut off sharing practices or even destroy value by making the shared data less valuable than the sum of its parts, while a well-functioning standard does the opposite and can become a foundation for broader sectoral cooperation. In the oil industry, ISO standard 15926 has for over a decade assisted firms in sharing lifecycle information about technical installations and their components. The standard works by converting information from a specific application to the standard format, after which any member of a project can use the information by importing it through their own adapter, adjusting it to their application needs.[23]

Testing the Limits: Three Important Constraints

Not all data should or will be shared. A fully articulated theory of data sharing would specify the conditions that define optimal boundaries between where shared data ends and non-shared data begins.[24] We are still far from advancing such a theory at this moment. Beyond the what, why, and how considerations that have made up the bulk of this paper's arguments, there are three additional and visible constraints that will shape theory and practice going forward. When should we simply exchange, rather than share data? How should personally identifiable information (PII) be treated when it comes to sharing? And how will political boundaries — in particular, the dynamics of national competitiveness that have become a significant factor in the data landscape— affect data sharing?

The first constraint is mostly about efficiency: will we discover conditions under which sharing *undermines* the efficient allocation of data to whoever can make the best use of it? Another way to pose that question is to ask, why doesn't the Coase theorem apply?[25]

In a Coase theorem world, with transaction costs low and property rights clearly defined, it does not matter who owns what data to start, because owners will simply make trades until data lands in the possession of the players who can generate the most value from it. This is an efficient outcome in the abstract. Data-sharing schemes would most likely undermine that efficiency by complicating property rights and (possibly) raising transaction costs. So if the data economy were actually operating in a Coasian world, the burden of proof should fall on data-sharing schemes to positively demonstrate that they create value above and beyond market exchange.

But data obviously does not live in a Coasian world. Transaction costs in practice are sometimes high, and property rights less than fully defined. Shared data could then be seen as a second-best outcome, an adaptation to those Coasian imperfections. Yet it is equally possible that data sharing gets in the way of Coasian dynamics and impedes the relative efficiency of market allocation, in which case it would be better to work on improving markets, rather than building schemes for sharing. The real question is, how do we know which of these approaches is right?

The answer is that, for many specific applications and settings, we do not yet know. In fact, some markets, such as online advertising, do seem to reflect a near-Coasian equilibrium, where there is less shared data (or visible demand for it) and more of a functioning marketplace where players buy and sell in a dynamic fashion. This is probably efficient because the property rights are clear, transaction costs are low, and it is reasonable to place measured values on the data being traded.

The point here is not to assume that existing arrangements are optimal or efficient just because they exist. A better approach is to view data-sharing schemes through the lens of a thought experiment, in which equal effort is dedicated to improving normal market exchange. The opposite thought experiment — replacing markets with data sharing schemes in places where markets dominate — is equally important. Real experiments would be even better than thought experiments. That's not simple, but it's the right approach over time.

The second constraint, a special case of data property rights, has to do with personally identifiable information (PII) and the general category of privacy concerns. Structurally, this is a meta-data problem, but not a trivial one. Even in non-shared datasets, privacy and related data protections are frequently controversial. The challenges have a way of multiplying when data is shared, and not only because the question of who becomes liable for a violation or breach is more complicated. More fundamental questions arise when discrete datasets with different

permissions or PII characteristics are combined. What permissions attach to shared datasets? And what about inference products that might pose greater privacy risks than any of the prior unshared raw data? That same question can and should be posed to market exchanges that bring datasets together, though the question is likely to appear earlier and more intensely for more extensive data-sharing schemes.

This constraint received considerable attention in the U.S. national security community following the 9-11 attacks, when the many separate agencies within intelligence and law enforcement were chastised for their lack of data sharing in advance of the attack. The 9-11 Commission argued that more data sharing might have led to better advance warning, while recognizing the deep sensitivities that Americans maintain about protecting citizens' perceived privacy from government intrusion.[26] The outmoded systems used by various government agencies had to some extent disguised this problem behind software barriers; many datasets that might have had shared intelligence value could not be easily shared because they were in old and incompatible formats. It seems far more desirable to make conscious, voluntary decisions about what should and should not be shared, rather than having the boundaries decided in effect by technology choices made years ago. But making such decisions requires confronting PII and privacy considerations squarely and candidly.

The third constraint on data sharing stems from political boundaries and emerging national competitiveness and national security considerations. Some first-tier national security arguments are obvious and have become salient as national defense organizations rely more heavily on data and machine learning applications. The Pentagon is not going to share relevant datasets (such as autonomous weapons performance data) with the PLA, and vice versa. Decisions about data sharing with allied defense establishments also are not straightforward.

Beyond these immediate military considerations, the landscape for data sharing across political boundaries becomes more complex and interesting. If data sharing is a source of value creation for leading sectors, then it will naturally become a locus of contention in national competitiveness debates — and not only between adversaries. Autonomous vehicle data is not only relevant to tanks, but is also a critical enabler for next-generation transport. If governments continue to see that sector as a core component of overall national competitiveness, the data becomes a strategic asset that might be limited to sharing within national borders and not outside. Similar dynamics could easily emerge in pharmaceuticals, machine language translation, or even supply chain optimization.

The surge in data localization laws that began during the late 2010s — in countries as different as Germany, Australia, Russia, and South Korea — is partly a function of this dynamic. Government policies that shape the sharing of data among private firms operating across national boundaries are becoming a manifest instrument of national competitiveness, before decision-makers have a clear theory about how and where it makes sense. When operating under such uncertainty, governments are by instinct more likely to err on the side of caution (i.e., limiting cross-border sharing) than they are to accept what seem like excessive risks in allowing data to cross borders. Rhetoric around AI/ML "arms races" exacerbates those instincts.[27]

In the longer term, the importance of national borders to debates over data sharing may recede. Data, after all, does not respect lines that divide countries on a map, and there are likely more efficient and effective ways of protecting national advantage than restricting data sharing. But in the contemporary global macroeconomic environment, national borders will likely constrain data-sharing schemes in surprising ways, which may cause positive reinforcement cycles of data nationalism to become more deeply entrenched. As with many global trade restrictions, governments will assume that if other governments are enacting protective barriers, there is something worth protecting on the other side. No one will want to be the last mover in that game, regardless of the deeper rationales at play.

Conclusion

Data sharing offers immense potential for various kinds of value creation in both the private and public sectors. There is also potential to destroy long-term value, to create more friction and transaction costs, and to reduce productive investment in the data economy. Whether positive or negative effects predominate depends on how data-sharing schemes are configured, structured, and explained. The key questions that need to be addressed are: what data is in play, what actors are involved on what terms, and to what goals and objectives is data sharing intended to contribute?

The evolution of the data economy is still in its early stages. Thoughtful and strategic bargaining mindsets are preferable to rigid ideologies and possibly inappropriate or dated notions of property rights. Incrementalism has important benefits. Data sharing is a new and unfamiliar practice for many organizations. Trying to go too fast is a recipe for problems, and firms would be wise to start with modest ambitions in relatively auspicious settings to

demonstrate proof of concept. A cautious approach is economically rational as it is difficult to foresee how much value can in practice be created or released. To try to avoid "no recourse" situations seems equally reasonable. Worries about ratchet effects (when once data is shared, it is always shared) are a constraining factor and could limit organizations' willingness to experiment.

Experimentation is the most promising way forward. This paper's contribution is to delineate conceptual schemes that name key characteristics that make for more precisely directed experimentation and measurement of results. This modest but important step makes possible assessments of what is working that are grounded in hypotheses, not assumptions or ideology, and should over time build toward more general theory. There is still a long way to go.

Material incentives, legal and economic structures, and technical infrastructures will strongly influence how the data sharing landscape evolves. Narratives are also important, as how we talk about data sharing will shape how organizations parse their risk-benefit decisions under conditions of high uncertainty. The conceptual schemes in this paper should help shape narratives about risks and benefits that move away from reflexive instincts about what is worth trying. This experimental approach will be a critical ingredient as institutions learn how to squeeze as much value as possible out of the evolving data economy.

Mini Case 1: A European Strategy for Data

In February 2020, the European Commission released a "European Strategy for Data," which aims to build toward a single market for data that would boost European competitiveness and set distinctive characteristics for privacy, governance, and access. A significant part of this strategy is a proposition for the EU to invest in a shared data infrastructure that would support common European data spaces in nine specific sectors. The strategy names the period between 2021-2027 as the window for investment in the technical and legal infrastructures needed to support data sharing in so-called strategic sectors and "domains of public interest," such as manufacturing, environment, mobility, financial, health, public administration data, and several others.[28]

The language of the Commission's proposal is expansive and ambitious, reflecting a strong desire to jump-start data-sharing schemes in what the Commission calls "non-personal industrial data and public data," with the expressed goal to drive growth, innovation, and improved decision-making. The plan goes so far as to say that "non-personal" data "should be available to all — whether public or private, big or small, startup or giants." This could be read as suggesting a dramatic point of view in which only personal data is legitimately enclosed or protected, though it is hard to imagine such regulation would require an oil company to share all its geologic data, make a homeowner share data on the performance of her solar panels, or demand that a biopharma share proprietary research data. Precisely where the lines between non-shared and shared data should be drawn, however, is not clear. The paper does argue explicitly that "data sharing between companies has not taken off at sufficient scale," blaming that failure on some combination of misaligned economic incentives, lack of "trust" in contracts, imbalances in negotiating power, lack of appropriate security, and legal ambiguity about who can do what with shared or co-created data — many of the ingredients of market failure that are generalized in data sharing. The document also notes how the lack of a "data-sharing culture" leads to sub-optimal business-to-government sharing, though it does not define what that cultural deficit looks like, where it comes from, or how it operates in practice. It does address in general terms some of the other limiting factors that impede data sharing,

including imbalances in market power, data interoperability and quality considerations, governance issues, and — importantly — the availability of infrastructure that can host, manage, and maintain shared datasets.

On balance, the Commission's proposal identifies key issues, even if it does not solve for them, marking a positive but partial step toward building workable schemes. The proposal charts a path forward that would build a framework for "common European data spaces" by legislating governance structures that prioritize standardization to enhance interoperability, ensure GDPR compliance, and provide a level playing field for small- and medium-sized enterprises. The strategy explicitly acknowledges a risk that shared data sets could enhance the market power of large incumbents who may be in the best position to extract value. It also recognizes the need to confront hurdles around intellectual property that data sometimes encodes, as well as legal liability issues.

The strategy has been criticized on a number of dimensions. One view is that it puts forward solutions in search of problems, by calling for the creation of personal data spaces without independent evidence that consumers actually want this. Another criticism has been aimed at the proposal for *de facto* data localization, as the strategy calls for cloud service providers that are owned and operated in Europe to host European data, with the implication that U.S.-based cloud providers are not secure, trustworthy, or ready to implement shared data schemes as the EU wishes. The ambitious plan to create common European data spaces rests on an untested assumption that there is market failure among private-sector players to provide these on their own. The strategy also includes implied conditions under which data sharing would appear to be forced by government action, above and beyond the narrow scope of public safety and related areas.

These criticisms are partly motivated by ideology and partly by different views of what data sharing is and is not, as well as by different assumptions about the costs, benefits, and risks of particular data-sharing schemes.

Mini Case 2: Data Sharing In Public Health Emergencies

2020 saw the launch of a large number of rapid scientific initiatives around the world to respond to and combat the COVID-19 pandemic. In the face of such public health emergencies, one of the most challenging yet critical components in building preparedness and an effective response lies in coordinating and combining sources of relevant data. Data sharing is widely recognized as fundamental to building more effective healthcare under normal circumstances, and even more so in emergency response systems, where time is of the essence. In recent years, we have seen a variety of multilateral efforts to address the needs and challenges involved. Lessons learned from the Ebola epidemic played a significant role in shaping some of the emerging efforts to support data sharing during the COVID-19 pandemic.

The experience of the Ebola epidemic was not positive. Early on, there were a number of individuals and organizations that were unwilling to share data in real time, including data that was considered mission-critical for public health planning.[29] At the time, experts indicated that data-sharing deficiencies were the result of a toxic mix of three factors: perceived disincentives or negative consequences for sharing data; the lack of a mechanism to enable data sharing; and insufficient positive incentives. For example, it was thought that data sharing could jeopardize journal publication; allow pre-emptive use of data by others for their own publications; and breach confidentiality agreements.[30] While some of the challenges in creating effective data sharing mechanisms remained unsolved, medical journals were in a particularly strong position to address some of these disincentives.

Two years into the Ebola outbreak, the International Committee of Medical Journal Editors (ICMJE), a working group of editors from leading medical journals around the world, agreed that pre-publication data sharing would *not* impact journal publication in the context of public health emergencies.[31] Subsequently, the ICMJE enacted a policy *requiring* that any data of critical importance to public health be shared with the World Health Organization (WHO) prior to publication. While these decisions seem like relatively easy and necessary steps toward removing disincentives for sharing research data, the WHO did not have in place a well-organized central data repository with public access. In fact, a large majority of the most

current datasets are exclusively available to WHO affiliates. Even with these developments, the Ebola pandemic was still seen as an indicator of the inadequacy of data-sharing arrangements to make material improvements in response to a public health crisis.

The COVID-19 pandemic has further emphasized the need for a central platform with robust access and near-real-time data. Although it is still too soon to fully evaluate the efficacy of the many efforts aimed at supporting data sharing in response to the pandemic, there are a few emerging efforts with strong institutional backing. In the private sector, HCA Healthcare, Google, and SADA have teamed up to launch the COVID-19 National Response Portal (NRP), an open-data platform built on Google Cloud technology and operated by SADA, a Google partner that provides technology services to the healthcare industry.[32] NRP enables U.S. hospitals to submit anonymized data to a central platform to provide a comprehensive real-time view of the COVID-19 pandemic. The portal includes metrics such as ICU bed supply and utilization; ventilator supply and utilization; and total numbers of positive, negative, and pending COVID-19 test results. The platform also incorporates publicly available datasets, including local shelter-in-place policies and mobility patterns that contribute to inferential models of how people and policy impact the spread of COVID-19.[33]

Another noteworthy emerging data-sharing platform is the COVID-19 Data Portal, which was created to accelerate coronavirus research to better understand the disease and develop treatments and vaccines. The portal launched in April 2020 as a joint effort between EU member states, the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI), and private research partners. The portal acts as a central data repository that enables researchers to upload, access, and analyze COVID-19 research data.[34] In addition to hosting relevant datasets submitted to EMBL-EBI, as well as other available biomedical datasets on sequences, protein structures, and other factors, the portal also provides data standardization guides and data analysis and computation tools.[35] Users can also access a list of other open data sets that are regularly updated with relevant information regarding COVID-19 research.

Public health emergencies have incentivized action from key players in a range of sectors, including industry, government, academia, and non-profit. These actors have tried to remove disincentives to sharing data in the context of public health emergencies; create positive incentive structures, by making some data sharing mandatory for journal publication; and build various types of infrastructure to enable real-time data sharing and access. While the responsive nature of these efforts may delay their ability to help with the acute phase of the 2020 crisis, they build capacity for more effective data sharing when the next emergency arises, and may contribute to experiments in data sharing under less exigent circumstances.

Mini Case 3: World Bank Open Data Initiative

Since adopting an Open Data Initiative mandate in 2010, the World Bank has served as a champion for expanding open data capabilities in governments. In addition to providing technical assistance or funding to data-sharing projects in over 50 countries, the World Bank has hosted over a dozen conferences on the role of data in government. The World Bank also freely offers all of its country research as open data, which it hosts on its Data Catalog platform. For data to be considered "open" under the World Bank, it must be fully available in the public domain, or accessible with "with minimal restrictions."[36] The World Bank also imposes technical standards on client countries looking to adopt open data policies, requiring that governments make their data available in machine-readable formats.

COUNTRY-LEVEL ENGAGEMENT: OPEN DATA INITIATIVES

As the world's largest development bank, the World Bank focuses much of its effort on funding and supporting capacity-building projects for client countries. These requests for open data usually come directly from governments themselves, and are tied to larger political campaigns involving improved transparency or accountability. For this reason, they are often led by senior political leaders or ministers within finance and tax departments.[37]

Upon agreeing to work with a client country, the World Bank walks the government through the launch of its own Open Data Initiative, using several methodologies and policy structures to assist in designing and diagnosing data sharing requirements. Often, a preliminary evaluation is conducted by an Open Data Readiness Assessment (ODRA) team, which helps governments evaluate interest and identify areas most in need of improvement. In Paraguay's ODRA, key recommendations included consolidating existing data catalogs and drafting binding policy commitments for senior leadership.[38] ODRA reports are also used to identify key stakeholders and champions needed to advance open data initiatives beyond World Bank engagement.

Following this initial assessment, governments establish short-term policy or institutional goals with help from the World Bank. As an early metric of success, countries will often look to set up an Open Data portal for its citizens as a benchmark toward future commitment. After receiving an action plan from its ODRA in 2014, Burkina Faso was able to pass legislation granting access to government data and establishing a Burkina Open Data Initiative (BODI) team. Through BODI, Burkina Faso was able to develop an Open Data portal containing over 200 datasets from 31 public-sector agencies, and create several web applications, providing real-time election results, water well identification, and geolocated school districts.[39]

In the process of helping client countries establish open data initiatives, the World Bank has also managed to identify several roadblocks to governments' long-term adoption of open data infrastructure. In a five-year report on lessons learned while implementing open data, World Bank officials noted that countries can find it difficult to maintain momentum once they have initially broken ground. Following the launch of Kenya's data portal in 2011, the government subsequently lagged behind in sustaining and expanding the platform beyond its initial launch.[40] Similarly, in several cases, administrations have found it difficult to justify sustained funding for data initiatives, since initial projects are often backed by external stakeholders, such as the World Bank or private donors. For open data projects that owe their existence to larger political campaigns, this can result in a short-term perspective that open data is a one-time project to be accomplished, rather than an undertaking to be continually funded. This project-based mindset also extends to civic engagement, with some ODRA teams noting a lack of government outreach and community response following the launch of open data initiatives.[41] In Paraguay, the open data platform established by the government lacked specific analytic tools to understand what data was being accessed, who was accessing it, and whether it was being used. Concerns also emerge surrounding equity, a data sharing problem not unique to the public sector. Some government officials have noted an unease with sharing data publicly, fearing either national security concerns or desiring a share of profits generated by private-sector usage.[42]

INTERNATIONAL ENGAGEMENT: RESEARCH AND DEVELOPMENT

Through its experience in country-specific projects, the World Bank has taken a leading role in conducting studies on the efficacy of open data in governance. In 2019, the International Development Association, an agency within the World Bank, published a report detailing specific data and statistical packages for client countries to use to improve evidence-based

policymaking.[43] The Global Governance Practice, another World Bank agency, performed a multi-year study on the impact of open data in the education and public-health sectors. The study established a strong positive association with open data and accountability in the education sector, and found it improved public-health management and health outcomes.[44]

The World Bank's commitment to open data has been an ongoing conversation. Prior to 2010, the World Bank locked much of its economic and development datasets behind paywalls. Following the emergence of open data, however, internal conversations within the World Bank pushed the organization to adopt policies that were more consistent with its mission to promote accountability and transparency.[45] Since then, all datasets from World Bank studies have been made freely available on the searchable Data Catalog platform, providing external researchers and other organizations with added analytic and visualization tools. The World Bank has also continued to assess its role in open data and explore new partnerships. The Independent Evaluation Group, a department that evaluates the development effectiveness of the World Bank, published a Data for Development report in 2018 highlighting the need for increased resources and coordination among teams.[46] Responding to the coronavirus pandemic, the World Bank also recently initiated the development of a COVID-19 Funding Tracker in collaboration with the Dutch government, allowing groups to assess humanitarian aid and outreach in different countries.[47] In the private sector, the World Bank has also been launching "development data partnerships" with large tech companies such as Google and Facebook, providing international organizations with new sources of data for evaluating public policy, such as social mobility or GPS data.[48] Looking ahead, the World Bank's 2021 annual World Development Report, titled "Data for Better Lives," indicates a continued investment in exploring open data solutions in governance.[49]

Mini Case 4: Shared Data Intermediaries

The last several years have seen a substantial evolution in thinking about the types of infrastructures needed to facilitate data sharing. Despite the existence of peer-to-peer (P2P) and direct file-sharing technologies, third-party platforms for hosting, accessing, and utilizing datasets between organizations were often not viable options. Mirroring the rise of AWS, recent years have brought increased development of cloud-based data storage and marketplaces designed specifically for companies to securely share and trade data. These industry-neutral "shared database intermediaries" provide infrastructure for organizations to host and securely exchange or share data with trusted parties.

A good example is Snowflake, a cloud-based data storage company that launched in 2012 and has grown rapidly into an emerging market leader. As of mid-2020, Snowflake had a valuation above $10 billion, having tripled its revenue and customer base for three years in a row.[50] In addition to its scalable data storage capacities, Snowflake's core value proposition is cloud-based data sharing, allowing customers to host and share data with other users without requiring them to maintain copies in their local environment. Snowflake also offers data-as-a-service (DaaS) marketplaces and data exchanges that let customers utilize or buy datasets from other companies. Once shared, these datasets can be used in conjunction with any accessible dataset, local or otherwise, to perform aggregate analytics or improve data quality.

As with cloud-computing services, Snowflake's role as a custodian and developer of infrastructure frees customers from having to divert capital expenditures to maintaining secondary platforms within their industry, and instead lets them access capacity on a flexible operating expenditure basis as and when they need it. This enables Snowflake in turn to prioritize gatekeeping tools for ease of use, allowing customers to disseminate data and delegate access in precise ways without sacrificing privacy or competitive information.

There are several types of incentives that might drive a company's decision to choose Snowflake for data-sharing purposes: 1) sharing data publicly for collaborative value or marketing exposure; 2) sharing data privately, as part of larger information exchange agreements between organizations; and 3) selling data to interested parties.

Public sharing: Public datasets can be used in collaboration with others to refine accuracy or address large-scale challenges. One of Snowflake's offerings is Data Marketplace, which permits approved organizations to be designated as "data providers," allowing them to publish standard, free-to-use datasets for other Snowflake customers. Data providers are also free to use data marketplace datasets to improve their own offerings. As a recent example, Weather Source, a weather and climate data analytics company, was able to improve the accuracy of a pandemic-related analysis by corroborating its research findings with public datasets from two other companies on Snowflake's Data Marketplace. It then published this newly curated dataset on Snowflake for researchers to use in linking seasonal factors to COVID-related infection rates.[51] By serving as a hub for datasets, shared data intermediaries are able to create an environment that encourages continuous improvement and cross-checking among customers. Public datasets are also used as a form of marketing exposure for generating interest and new connections. Because Snowflake datasets are organized by industry, data providers can post datasets to be "test-driven" by prospective customers, much like the "freemium" model frequently used to market applications.

Private Sharing: Shared data intermediaries can also allow for more confidential transactions between organizations while reducing overhead costs. For industries that include multiple third-party contractors or shared resources, it can be difficult to coordinate data sharing without creating duplicative (and therefore stale) data. Snowflake allows customers to create private data exchanges for approved users within and across organizations. This enables organizations to avoid redundant copying of data among partners, and is built specifically with scalable data sharing and privacy controls in mind, ensuring that competitive data is not inadvertently leaked.[52] In the upstream oil and gas industry, for example, it is common practice for well operators to share ownership due to production and operational costs. To coordinate usage and progress reports, data sharing is a core requirement between partners to provide drilling operation and project updates. Previously, operators had to maintain their own customized platforms, disseminating relevant data to partners through file transfers or emails. This method often resulted in redundant IT platforms created by technology consultancies, which required additional development and increased maintenance costs over time.[53] Snowflake has been able to sidestep these problems for several oil and gas companies, including Devon Energy and Uniper Energy, enabling them to reduce data warehousing costs, eliminate data silos, and improve controls for user access.[54]

Data Selling: Whether considering data as a primary product, or seeking to monetize a pre-existing dataset collected for other purposes, organizations sometimes want simply to sell their

data as a means of generating income. Snowflake enables data providers to offer premium datasets for sale on the Data Marketplace, and allows them to customize the dataset for specific customers. Data providers are given options for several types of payment structures, including one-time purchases or subscription pricing. In opening up DaaS products to other buyers, organizations can increase the perception of value in their own offerings while controlling how much and to whom access is granted.[55]

The market for data-as-a-service offerings reached about $26 billion in 2019, and is expected to double in the next five years.[56] As the value of data increases, the channels in which it can be found, shared, and sold will need to evolve accordingly. Shared data intermediaries play an important role in this marketplace. The variety of offerings they create should continue to draw in new customers and extend data-sharing practices and experiments.

Mini Case 5: Data Sharing Experiments in the Pharmaceutical Sector

The pharmaceutical industry has for several decades been searching for needed improvements in research productivity. At roughly 15% of revenue, Pharma's R+D investment is among the largest of any major economic sector.[57] Despite that investment, financial returns on R+D spending have been on a downtrend and are lower than in many other sectors. Compared to 2010, when there was a 10.1% return on R+D spending, 2019 saw just a 1.8% return.[58] Other measures of productivity, such as the development of new molecular entities (NMEs) approved for use, have shown similar declines. This productivity challenge has catalyzed major pharmaceutical companies to experiment with how they use and share data, particularly in the pre-competitive domain, where the boundaries are expanding to include clinical research.[59] Clinical trial data has become the focal point of data-sharing efforts in the pharmaceutical sector because there is great potential to advance science, improve patient outcomes, and drive commercial success at once.

For clinical trial participants, sharing data in theory would increase their individual contributions to generalizable knowledge about science and medicine by potentially facilitating additional findings beyond the predetermined clinical trial purposes, which is an additional incentive to join and remain in clinical trial protocols.[60] Additionally, sharing clinical trial data could reduce unnecessary duplicative effort and costs of future studies by different research groups; minimize exposure of participants in future trials to avoidable harms identified in previous trials; and develop a more robust database for researchers, regulators, and patients.[61] For these reasons, the United States and Europe require that clinical trials be registered to ClinicalTrials.gov and the EU Clinical Trials Register respectively, within timelines ranging from six months to one year after the end of the trial.[62] The trials must also submit "basic results," which by the Food and Drug Administration's (FDA) definition include: participant flows, baseline characteristics, outcome measures and statistical analyses, and adverse events.[63] These results are reported in summary format and do not include patient-level information, so should be seen as an early and quite limited (though still notable for the industry) form of data sharing.

Data sharing does create risks and concerns for both the patients and the company sponsors. At the patient-level, the central concern is maintaining privacy.[64] There are well-established mechanisms to address de-identification needs, including standardizing and anonymizing how patient-level data is handled and shared — a practice that researchers and organizations also exercise when publishing clinical research in peer-reviewed journals. These mechanisms are imperfect, and trade-offs with the ability to access scientific value from data are present but not fully measurable, which in turn makes strategy and decision-making around data sharing more difficult.

The calculations about sharing data are not straightforward for firms, either. For drug discovery, data is of course the life-blood of the industry and is often the primary source of competitive advantage. While data sharing may be easier to imagine and implement in the precompetitive domain, the boundaries of that domain within the pharmaceutical industry are expanding as technology, business models, and the science of drug discovery change.[65] Introducing data sharing at the clinical research stage could shift competitive advantage in unpredictable ways. For instance, confidential and proprietary information could be discerned from the data, or in other cases, competitors may exploit shared data to make and sell competing products in countries that do not protect the original research.[66] While these risks are commonly understood within Pharma, the benefits are more challenging to demonstrate, as the full utility of the data is unknown before it is shared. Due in part to these factors, legal and IP departments within pharmaceutical companies (which are generally responsible for deciding if data can be shared) have tended to downplay the potential benefits, and by default tend to resist non-mandated data sharing.[67]

Still, stakeholders inside and outside the firms have recognized data sharing's potential to create collective benefits, and are experimenting with creative solutions to enable sharing. In an effort to mitigate risks and improve R+D productivity, major pharmaceutical companies like Pfizer[68] and Sanofi[69] are teaming up with academia, government, and smaller biotech firms to participate in independent platforms that aim to facilitate and accelerate the drug discovery process. Vivli, for example, is an independent non-profit, cloud-based global data sharing and analytics platform designed to make clinical trial data accessible for researchers worldwide. This organization evolved from a project of The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard (MRCT Center), and aims to enhance access to clinical trials data by promoting data sharing and transparency.[70] In addition to acting as an independent data repository, Vivli provides a secure research environment within Microsoft Azure Cloud that includes access to analytical tools for researchers. The platform coordinates

and integrates existing data-sharing initiatives and includes more than 5100 studies, with contributions from about 3 million trial participants from 115 countries — a significant expansion from what was possible or available 10 years ago.[71]

Focused exclusively on oncology, Project Data Sphere is another example of an independent platform that pairs an open-access data-sharing model with special research tools.[72] The Project has relationships with government entities in the U.S. and Europe, and industry partnerships that span some of the largest pharmaceutical companies. While platforms like Project Data Sphere and Vivli aspire to help companies improve their returns on investments by streamlining the R+D process, the goal ultimately is to improve the quality of care for patients. These initiatives' success in achieving these goals will rely on continued collaboration and contributions from industry and regulators.

Endnotes

1 Lisa Jardine, *Ingenious Pursuits: Building the Scientific Revolution*. New York: Anchor, 2000. 17–19.

2 Comprehensive review is in Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

3 Steven Weber, "Data, Growth, and Development," *Business and Politics* 2017, 19(3), 397–423.

4 For example at https://www.sciencedaily.com/releases/2013/05/130522085217.htm.

5 Steven Weber, *Bloc By Bloc: How to Build a Global Enterprise for the New Regional Order*. Cambridge MA: Harvard University Press, 2019.

6 Weber, *Bloc by Bloc*.

7 Thomas S. Kuhn, *The Structure of Scientific Revolutions,* Chicago: University of Chicago Press, 1962.

8 For example see Andrew Vickers, "Sharing raw data from clinical trials: what progress since we first asked 'Whose data set is it anyway?'" *Trials* 2016 17:227.

9 See for example Bowman Heiden and Nicolas Petit, "Patent Trespass and the Royalty Gap: Exploring the Nature and Impact of Patent Holdout", *Santa Clara High Tech* 34 179 (2018), https://ssrn.com/abstract=2981577.

10 The agreement could be a strong implicit one or an explicit one. The lack of any such agreement suggests something other than data sharing per se.

11 Weber, "Data, Development and Growth".

12 This could include the absence of any obligation with regard to additional data that the ML system produces as it functions in the wild.

13 Steven Weber, *The Success of Open Source*. Cambridge MA: Harvard University Press, 2004. Note that definition of 'use' can vary but in traditional GPL licensing generally means "compiles with."

14 Specific reciprocity implies that the ledger balances around individual transactions as they occur, while Diffuse reciprocity implies a general balancing over time.

15 FRAND is an acronym for Fair, Reasonable, and Non-Discriminatory Licensing. The European Commission provides a good overview from 2015 at https://publications.jrc.ec.europa.eu/repository/bitstream/JRC96258/jrc96258.pdf.

16 Such as the slogan "information wants to be free," generally attributed to Stewart Brand and/or Steve Wozniak, which was an early expression of non-rivalness.

17 Mancur Olson, *The Logic of Collective Action: Pubic Goods and the Theory of Groups*. Cambridge MA: Harvard University Press, 1971.

18 Many of these issues are reviewed in Chris Johnson et. al., *Guide to Cyber Threat Information Sharing*, NIST Special Publication 800-150, US Department of Commerce 2106, at https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-150.pdf.

19 Ronald Burt, *Structural Holes: The Social Structure of Competition*. Cambridge MA: Harvard University Press, 1995.

20 This is the hegemonic or K-group (privileged group) solution from Mancur Olson's work.

21 Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions. *A European Strategy for Data* 19 February 2020. At https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf See in particular the 10 specific 'Common European Data Spaces' proposal ideas in the Appendix to the report.

22 Adversarial machine learning attacks are the most prominent example of an intentional exploit in this manner.

23 ISO technical description at https://www.iso.org/obp/ui/#iso:std:iso:15926:-13:ed-1:v1:en.

24 Like transaction cost theories of the firm, which specify the boundary between firm and market. As with TCE, the theory could be descriptive or prescriptive/normative.

25 Ronald Coase, "The Problem of Social Cost," *Journal of Law and Economics* 1960 3 (1) 1–44.

26 National Commission on Terrorist Attacks, *The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks Upon the United States.* New York: WW Norton, 2004.

27 Steven Weber, *Bloc by Bloc*, Chapter 6.

28 Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions, *A European Strategy for Data*, 2 Feb 2020, at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN

29 https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)00758-8/fulltext.

30 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7136984/.

31 https://www.who.int/bulletin/volumes/98/3/20-251561/en/.

32 https://nationalresponseportal.com/about/.

33 https://investor.hcahealthcare.com/news/news-details/2020/HCA-Healthcare-Teams-With-Google-Cloud-and-SADA-on-Data-Portal-to-Help-Communities-Respond-to-COVID-19/default.aspx.

34 https://www.covid19dataportal.org/.

35 https://www.covid19dataportal.org/about.

36 http://opendatatoolkit.worldbank.org/en/essentials.html.

37 Interview with Craig Hammer, World Bank Program Manager

38 https://documents.worldbank.org/en/publication/documents-reports/documentdetail/711141530859805700/open-data-readiness-assessment-government-of-the-republic-of-paraguay.

39 https://schoolofdata.org/2016/11/05/the-state-of-open-data-in-burkina-faso/.

40 https://openknowledge.worldbank.org/bitstream/handle/10986/28616/120801-WP-P133276-PUBLIC.pdf?sequence=1&isAllowed=y.

41 Interview with Silvana Kostenbaum, ODRA Public Sector Specialist

42 https://openknowledge.worldbank.org/bitstream/handle/10986/28616/120801-WP-P133276-PUBLIC.pdf?sequence=1&isAllowed=y.

43 http://documents1.worldbank.org/curated/en/696731563778743629/pdf/IDA19-Second-Replenishment-Meeting-Special-Theme-Governance-and-Institutions.pdf.

44 http://documents1.worldbank.org/curated/en/410191559670657041/pdf/From-Theory-to-Practice-Open-Government-Data-Accountability-and-Service-Delivery.pdf.

45 Interview with Craig Hammer, World Bank Program Manager

46 https://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/datafordevelopment.pdf

47 https://blogs.worldbank.org/opendata/new-tool-facilitates-greater-transparency-and-coordination-covid-19-financing.

48 https://datapartnership.org/updates/.

49 http://documents1.worldbank.org/curated/en/778921588767120094/pdf/World-Development-Report-2021-Data-for-Better-Lives-Concept-Note.pdf.

50 https://www.prnewswire.com/news-releases/snowflake-more-than-triples-revenue-and-customer-base-doubles-post-money-valuation-all-in-just-one-year-300793857.html, https://www.ig.com/en/trading-strategies/snowflake-ipo-200630,, https://www.forbes.com/sites/petercohan/2019/08/16/growing-at-237-snowflake-says-its-taking-business-from-teradata-and-ibm/#5383d72d2186.

51 https://weathersource.com/blog/weather-source-publishes-covid-19-dataset-on-snowflake-data-exchange/.

52 https://www.sbconsulting.com/blog-category/simplifying-oil-and-gas-data-sharing-with-snowflake.

53 https://www.drillingcontractor.org/a-broken-model-data-ownership-sharing-control-still-await-solutions-44130.

54 https://polestarllp.com/little-book-of-big-success-cloud-data-platform-edition.pdf, https://www.businesschief.eu/company/how-uniper-powering-its-digital-transformation-cutting-edge-data-analytics.

55 https://towardsdatascience.com/how-to-sell-your-dataset-2b4581175a738.

56 https://www.mordorintelligence.com/industry-reports/data-as-a-service-market.

57 https://www.statista.com/statistics/309466/global-r-and-d-expenditure-for-pharmaceuticals/

58 https://www2.deloitte.com/uk/en/pages/life-sciences-and-healthcare/articles/measuring-return-from-pharmaceutical-innovation.html#

59 https://www.ncbi.nlm.nih.gov/books/NBK54325/

60 https://www.ncbi.nlm.nih.gov/books/NBK285999/#:~:text=In%20the%20long%20run%2C%20sharing,or%20less%20effective%20than%20alternatives.

61 https://www.ncbi.nlm.nih.gov/books/NBK285999/#:~:text=In%20the%20long%20run%2C%20sharing,or%20less%20effective%20than%20alternatives.

62 https://www.pfizer.com/science/clinical-trials/trial-data-and-results#:~:text=CLINICAL%20TRIAL%20DATA%20SHOULD%20BE%20ACCESSIBLE%20AND%20TRANSPARENT&text=Pfizer%20publicly%20shares%20results%2-ofrom,volunteers%2C%20researchers%2C%20and%20others.

63 https://clinicaltrials.gov/ct2/about-site/results.

64 https://www.ncbi.nlm.nih.gov/books/NBK285999/#:~:text=In%20the%20long%20run%2C%20sharing,or%20less%20effective%20than%20alternatives.

65 https://www.ncbi.nlm.nih.gov/books/NBK54320/

66 https://www.ncbi.nlm.nih.gov/books/NBK285999/#:~:text=In%20the%20long%20run%2C%20sharing,or%20less%20effective%20than%20alternatives.

67 https://thejournalofmhealth.com/data-sharing-how-pharma-can-benefit/

68 https://www.pfizer.com/science/clinical-trials/trial-data-and-results#:~:text=CLINICAL%20TRIAL%20DATA%20SHOULD%20BE%20ACCESSIBLE%20AND%20TRANSPARENT&text=Pfizer%20publicly%20shares%20results%2-ofrom,volunteers%2C%20researchers%2C%20and%20others.

69 https://www.sanofi.com/en/science-and-innovation/clinical-trials-and-results/our-data-sharing-commitments.

70 https://vivli.org/about/overview-2/.

71 https://vivli.org/resources/platform_metrics-2-2/.

72 https://www.projectdatasphere.org/about.

Acknowledgments

This project is made possible by a grant from the William and Flora Hewlett Foundation with additional funding from Facebook in support of independent academic research.

About the Authors

Steven Weber is Professor and Associate Dean at the School of Information and Department of Political Science at UC Berkeley; and Faculty Director of the Center for Long Term Cybersecurity (CLTC). His most recent book is *Bloc by Bloc: How to Organize a Global Enterprise for the New Regional Order* (Harvard University Press 2019)

Matthew Nagamine is the Manager of Strategic Partnerships for the Center for Long-Term Cybersecurity, where he manages the CLTC Corporate Membership Program and administers the CLTC Grant Program, among other roles. He earned his B.A. in Political Science and African American Studies from UC Berkeley.

Max Ingraham-Rakatansky is a Graduate Student Researcher at the Center for Long-Term Cybersecurity and Vice President of the Information Management Student Association at UC Berkeley. He is currently studying the impacts of online communities on social behavior at the School of Information.

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

Center for Long-Term Cybersecurity
cltc.berkeley.edu
@CLTCBerkeley