Harvard Data Science Review • Special Issue 4: Democratizing Data

A Mapping Lens for Estimating Data Value

Abhishek Nagaraj[1,2]

[1]University of California Berkeley, Berkeley, California, United States of America,
[2]National Bureau of Economic Research, United States of America

Published on: Apr 02, 2024

DOI: https://doi.org/10.1162/99608f92.82f0de5a

ABSTRACT

Public data is a key resource for society. The digital revolution and a renewed focus on evidence-based policymaking have made estimating the value of public data resources more urgent than ever. However, conventional methods to assess their value are still underdeveloped and generally fail to capture their broader impact. We introduce a novel approach to assess the value of data, anchored in its potential to improve decision-making. This approach likens data to a map, in that it provides an (imperfect) representation of a conceptual landscape that guides how people navigate the landscape. We define the value of data as the divergence in the quality of decisions taken with this map as compared with the closest alternative, and argue for the use of natural experiments to estimate this value. We demonstrate the use of this approach with four concrete examples drawn from our past research. Crucially, our method can inform how federal agencies decide to collect, maintain, and publish their critical data assets amid growing privacy concerns.

Keywords: public data, data value, quasi-experimental methods, government data

1. Introduction

Data is the lifeblood of decision-making in the digital age. As information that has been numerically coded and thereby made more amenable to analysis, it holds significant value for academic research (e.g., testing conjectures and generating scientific theories), government policy (e.g., defining and evaluating economic and social policies), and business and society (e.g., guiding management and investment decisions and consumer choices). For instance, researchers increasingly rely on publicly available, large-scale, centralized data sets to discover new insights in research fields as diverse as health, genomics, biology, climate change, ecology, economics, meteorology, and astronomy. Governments increasingly use sophisticated forms of data to target and evaluate policies on trade, labor and housing markets, taxation, and law enforcement. In commerce, the use of data-driven applications (e.g., location-based weather and transportation information) is now firmly integrated into consumers' everyday decision-making. Finally, firms in venture capital, real estate, pharmaceuticals, retail, and related industries use data to identify promising new investments and projects in the hopes of increasing their return on investment.

This article focuses on public data, which is data collected, funded, or made public by government institutions.[1] This includes so-called administrative data, which is nonstatistical information that is routinely gathered and stored by government agencies as a part of their everyday operations. This data is not collected to intentionally aid academic research or the private sector. Examples include unemployment insurance claims, Medicare data, tax records, patent applications, geo-satellite imagery, and so on. As we shall see, government agencies will often only make such information available to the public under strong restrictions. In addition, some agencies, such as the U.S. Census Bureau, are mandated to collect statistical information about the nation's demographic

and economic trends via surveys and census enumerations. Other agencies, like the United States Geological Survey (USGS), finance and gather geospatial data (e.g., satellite imagery), which is subsequently used for weather forecasts, planning transport networks, land cover, and so on. Finally, agencies like the National Institutes of Health fund scientific data collection to explicitly support empirical research and help test and generate scientific theories; noteworthy examples include the Human Genome Project, astronomical surveys, and protein databases. There is growing interest in assessing the economic value of such public data, including by the U.S. government itself. For instance, in 2018, Congress passed a transformative step in evidence-informed governance: the landmark Foundations for Evidence-Based Policymaking Act (2019; hereafter Evidence Act), on the explicit principle that the federal government should incorporate data and evidence-building into every step of its decision-making process (Potok, 2024). As part of the legislation, federal agencies are encouraged moving forward to make any data collected publicly available. This is because, very often, the value of data extends beyond its initial use cases inside the agencies. Indeed, as we shall see, businesses, not-for-profit organizations, academic researchers, and even individuals (i.e., the 'public') have also discovered the value of public data for their own decision-making. Of course, there are limits to what government agencies can publicly release, especially in light of growing confidentiality concerns. To navigate this new terrain, the Evidence Act established an Advisory Committee on Data for Evidence Building (ACDEB), which recommends the adoption of a 'risk-utility' framework that relates the privacy costs ('risk') of data assets with the benefits ('utility') of expanding access. While significant attention has already been paid to quantifying the risk of public data—for instance, with the development of a 'differential privacy' approach[2]—there is still an urgent need for agencies to develop better dollar measures of utility, which will require understanding the full economic value of their data assets. If such a framework existed, as it does for other public resources like public research funding, then we would be better able to make decisions related to the funding and maintenance of public data.

In addition, there are often operational expenses incurred when collecting and maintaining public data, which government agencies need to appropriately budget and account for. Indeed, these agencies may need to decide whether (or when) the expenses are justified in the first place, or conversely, whether there is merit to *increasing* data provision and pricing the resulting 'data products' to match the added economic value. Moreover, even if the value of one data set may exceed its cost, agencies often have limited budgets and must choose among different data sets to identify the one with the most value. Having accurate estimates of the value of public data can help to inform agencies as they attempt to navigate these difficult decisions. Lastly, government institutions often need to measure the quality of their service to clients, and currently rely on user requests (i.e., customer feedback) to do so. Yet we contend that (point-in-time) estimates of the economic value of their assets is arguably a more meaningful measure, and building a time series of these estimates would enable managers to measure progress over time. Anecdotally, our communications with current government practitioners suggest that these measures are much needed, as administrative data is often managed or shared sub-optimally.

However, inferring the value of public data can be somewhat challenging in and of itself. We cover the full set of challenges in detail in a later section, but for now we briefly summarize them here. First, as it stands, it is difficult to trace who uses public data, how, and for what purpose, particularly in many contexts where citations to data are less common. Second, data can spark a chain reaction in decision-making, which means that many users may not even realize that their decisions are being informed by such data. Third, there are myriad applications and a diverse user base, which can make it difficult to capture the totality of the data's broader impact. Fourth, there are always counterfactual scenarios—that is, what *would have happened* without the data—that are fundamental to data value but hard to define and measure.

Yet, despite these complications, there have also been a series of advancements in empirical methods in the social sciences that offer promising avenues for better assessments of data value. One such development is the use of 'natural experiments' (or 'quasi-experiments'). These are situations when nature produces experimental-like conditions, such that some subjects are exposed to a treatment, while others are not. In our context, if we can identify scenarios where a group of decision-makers have (coincidentally) been given access to a new or improved data set (we will call this the treatment group), and conversely a group that has not (the control group), then we can credibly quantify how informational differences can influence a decision-maker's choice of action. By comparing the quality of decisions taken when data was made available to those taken when it was not, we can overcome many of the challenges associated with ascribing value to data.[3]

Hence, this article will focus on quasi-experimental methods and their application to the valuation of public data. We provide a framework for data valuation that combines the use of these natural experiments with past work in cartographic theory. Our primary contention is that data essentially acts like a cartographic map, in that it provides an (imperfect) representation of a conceptual landscape that guides how people navigate the landscape. If we can find natural experiments that vary the quality of the map, we can infer the value of data by comparing the people's decisions under alternate representations of the same landscape. To fix ideas, this technique is analogous to valuing a traffic map based on differences in travel time when real-time information is unavailable (or suffers from network latency issues) in some regions but not in others due to random technical errors.

The framework is described in Section 2, while Section 3 provides a selected overview of research that uses this mapping framework to value public data. Section 4 outlines the advantages and disadvantages of alternative approaches to valuing data. Finally, Section 5 concludes by discussing challenges and future opportunities.

2. Data as a Map for Decision-Making

2.1. Challenges in Existing Methods of Valuing Data

Valuing public data can be a complex problem. This is primarily because, as it currently stands, it is not always clear *who* uses data, *when* they use it, and *for what* purpose. For example, tools like Google Dataset Search do not clarify which data set users find is the primary source versus derived versions, so it is hard to record the original data source (Sostek et al., 2024). Except for a few rare use cases, data users do not generally cite their usage or leave behind some other public record of their adoption, especially in individual or commercial contexts.[4] For instance, measuring the value that a retail business would derive from using census data to identify candidate store locations would not be trivial if there are no public records of them using census data in the first place. This is not unlike the challenge economists face when valuing internet-based technological innovations, such as social media and maps. These innovations are freely accessed with no explicit exchange of goods and services, which means their value needs to be imputed in gross domestic product (GDP) calculations.[5]

A second challenge is that individuals or institutions may not necessarily be aware that some data set is actively influencing their decisions. Consider an investment firm that, with some success, uses data-driven analyses to identify profitable investments. If its competitors begin to imitate these strategies, then the value of the data should encompass the benefits accrued to the lead investor *and* its followers, regardless of whether these imitators directly used the data themselves, or even knew it existed. This phenomenon tends to occur more generally in research and development (R&D), an area where firms have become particularly data-driven (see Tranchero, 2023, for one example in the pharmaceutical industry) yet are often vulnerable to a business stealing effect (Bloom et al., 2013). It should already be clear, then, that more robust empirical methods are needed to answer the question ‘What is the monetary value of public data?’ beyond simply counting the number of data downloads and requests; these figures likely represent the ‘tip of the iceberg.’

Third, it can be prohibitively challenging to quantify the value of data empirically, especially in light of the wide range of possible applications. This complexity is exemplified by the work of Hausen and Azarbonyad (2024), who developed a framework for extracting government data set usage from scientific publications using natural language processing techniques. While this represents a significant advancement, the varying importance and impact of these data applications on different decision-makers and the variety of potential users (e.g., governments, businesses, researchers, not-for-profits, and individuals) further complicate the valuation process. For instance, publicly available satellite imagery has been used by farmers to predict crop yields, by meteorologists to forecast weather events, and by city planners to improve urban design.[6] Typically, it may also be unclear precisely *when* the data contributes to a decision or *which subset* of the data does so, as these factors are rarely tracked. Moreover, for each of these applications, any estimated data value would (at best) be a snapshot of one specific point in time. Many types of data become more valuable over time, particularly as the userbase grows in number, or data creators introduce updates to improve comprehensiveness and granularity; this process could look very different depending on the type of application.

Finally, conceptually speaking, the value of data does not only depend on the decisions taken *in light of* such data, but also on the decisions that *would have been* taken had the data not been available (i.e., the 'counterfactual case'). For instance, even if a specific public data set is unavailable, it may still be possible for decision-makers to access a reasonably close substitute. Alternatively, their intuition may already align with the recommendations yielded by formal data analysis. In these cases, the *marginal* value of the data would be small. Unfortunately, such counterfactuals are hard to define, which is partly responsible for the current scarcity of empirical studies on the value of public data. Where estimates do exist, they tend to focus on narrow measures of value that are derived from direct usage or survey estimates, which (as we explain in greater detail in Section 4) are likely to be unreliable. As one recent exception to this, we highlight the burgeoning practitioner-led literature on the science of the value of information (VOI), which has produced estimates of data value informed by counterfactual decisions in a variety of environmental contexts, from satellite constellations to wildfire management (Lauer et al., 2021; Laxminarayan & Macauley, 2012; Simon et al., 2022).

While a holistic and robust accounting of the value of data can therefore be challenging, it is not impossible. Indeed, a suite of promising methodological developments in the empirical social sciences could enable more accurate estimates of the economic value of data-driven decisions. In particular, we highlight the increasing popularity of natural experiments as a research method — an innovation that recently won David Card, Joshua Angrist, and Guido Imbens the 2021 Nobel Prize in Economic Sciences (The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel, 2021). Natural experiments (otherwise known as quasi-experiments) attempt to mimic the rigor of clinical trials in medical research by discovering real-life scenarios that, by an accident of nature, randomly assign individuals into two groups—a treatment group and a control group—and comparing outcomes across them. In a famous early example, Card (1990) leverages the Mariel boatlift of 1980 to estimate the effect of immigration on wages and unemployment. They do so by comparing labor market outcomes in Miami (a city 'treated' with a sudden influx of more than 120,000 immigrants after the boatlift) to other comparable cities with no such immigration. Similar methods have been used to recover the returns to education (Angrist, 1990), the consequences of minimum wage law (Card & Krueger, 1994), and the income effect on labor supply (Imbens et al., 1999). Beyond economics, natural experiments have become ubiquitous (and highly accepted) in peer-reviewed academic research, including in political science, sociology, epidemiology, and other social sciences. As we shall see, we can also use these methods to recover estimates of public data value.

2.2. Maps and Their Relevance to Data Valuation

While the use of natural experiments looks promising, the next question is how we can apply these methods in the context of data valuation and identify what exposure to a treatment may look like. As it turns out, we can turn to cartographic theory (i.e., the study of making and using maps) for some remarkably useful insights into

these questions. To understand why, it is first necessary to consider what a map is in the first place, as well as some of its relevant properties.

In this article, we consider a map to be a stylized representation of a geographic data set.[7] Because other methods of conveying the same underlying information (e.g., coordinates with latitude and longitude) can be quite cumbersome to interpret—and maps are generally more intuitive and do not require specialized training—maps have become the standard way of conveying the location of geographic features. Of course, knowing the location of a river or rock formation is not necessarily useful in and of itself. Rather, maps have become widely embraced because they help humans *make decisions.* The most obvious decisions are navigational in kind. Maps helps an individual get from place A to place B. But, in the analog era, paper maps were also essential for street, housing, military, and urban planning. Nowadays, digital maps are frequently used in consumer-facing applications, such as real-time transportation, traffic, and weather information, or ridesharing apps, real estate portals, and local search engines.[8] Fundamentally, then, maps are tools to guide decision-making (such as which route to travel, which house to buy, or which destination to visit) and the value of the map is the improved utility that the decision-maker derives. For instance, this value could stem from reduced travel times from using Google Maps, or higher revenues from using satellite maps to target tax collection efforts (Casaburi & Troiano, 2016).

While maps can provide significant value, it is important to remember the foremost insight from cartography that "a map is not the territory" (Korzybski, 1933, p. 750). In other words, any given map is fundamentally a single representation of a physical space that contains (often useful) distortions, making it distinct from the ground truth. The possible distortions are numerous. Maps may differ in their 'resolution,' representing similar spaces at different levels of granularity, or in their coverage, ignoring certain parts of a landscape. Often, a given part of the terrain can be associated with many different characteristics; some maps may specialize in depicting a very specific set of features (e.g., building height), while others may be general and include an array of useful information (e.g., tourist sites). Irrespective of the representational choices they make, the fact remains that maps are inherently constrained and require deliberate choices about what to include and what to leave out. Further, these choices are rarely random happenstance. What precisely is represented and what is left out is often a function of systematic factors, like the incentives of mapmakers, customer demand and competition, intellectual property, and legal considerations (Nagaraj & Stern, 2020).

Crucially, these very distortions can provide us with the means to infer the economic value of the maps themselves (as well as some clues for how to value data more generally). Consider the following stylized example: there are two official maps that describe the same mountainous region (each with a distinct choice of representation), and one group of mountaineers is using the first map, while another otherwise comparable group of mountaineers is using the second. If there are material differences in the ability of these two groups to navigate the terrain, then we can attribute these differences to the specific representations and thereby impute the relative value of each map. Perhaps the two groups are engaged in a competition and the first to complete

the circuit will win prize money (which would make our job of estimating the value of the better map much easier). Or, each group simply values a quicker journey (which would require the additional step of valuing time). Nevertheless, as long as there is some outcome of interest, we can ascribe some measure of value.

With this in mind, we can begin to understand where to look for natural experiments. Per the example above, if we can find a group of decision-makers that use some map whose value we are interested in, we can designate this the treatment group. If we have a group of decision-makers that use the next-best alternative map (which has a different set of distortions), we can designate this the control group. This is not the only possibility. Consider a second instructive example, which (arguably) has more real-life parallels. In this example, there is only one official map, but one of the mountainous regions is represented at a higher fidelity, while another roughly comparable mountainous region is represented at a lower fidelity. If this creates some divergence in the ease of navigating the two regions (or more generally, the ability to make a decision)—which further downstream creates a divergence in travel times (or more generally, some outcome of interest)—then we can use the differences in fidelity to infer the value of a higher resolution mapping. Once more, we have a clear treatment and control group. Note that for this example, the two regions must be *otherwise comparable* or even identical, which in practice may present some difficulties since they are, indeed, different regions.

Of course, we are not solely interested in the value of maps, even if they do constitute one potential (and important) type of data. Rather, our ultimate objective is to value all forms of public data. Nevertheless, we can now make our central connection: every data set is kind of like a map. It is deliberately collected to describe some ‘landscape.’ A brief example here may be helpful. Pharmaceutical companies often spend significant time and R&D budget looking for the genetic mutations responsible for genetic diseases to guide their drug development (Tranchero, 2023). This can be a daunting task: there are tens of thousands of known human diseases, and any one of the thousands of human genes could be a potential candidate. Testing each possible gene-disease combination is logistically impossible, so it is difficult to traverse the ‘gene-disease’ landscape. Hence, in recent years, academics have introduced large public data sets that detect correlations between genes and diseases from medical studies—a useful first pass. In other words, this data gives pharmaceutical companies a map to navigate the gene-disease space. This equivalency means that the preceding discussion, and all the relevant properties of maps, also apply. This allows us to combine the insights from cartographic theory with the methodological ‘revolution’ in applied economics to develop a versatile framework that can recover estimates of public data value in a variety of settings. We introduce and build on this framework in the next section.

2.3. The Framework: A Mapping Lens to Value Data

The first principle of this framework is that data has no inherent value. This is a central tenet of decision theory: the value of data is completely derived from its use in decision-making. For example, knowing whether or not it will rain on any given day is useful insofar as it influences the decision to bring an umbrella to work. This information has little value if one always carries an umbrella, or conversely never does so. Other

examples where data might play a role in decision-making include a government deciding which subpopulation to target for an intervention, or an individual deciding which mode of transport to use for their commute. In both cases, the 'right' answer is unknown from the outset and will require consideration of the sundry options (each with uncertain payoffs), which data can help to illuminate. Therefore, our first principle is that the value of public data must be rooted not in its usage but in terms of the improved outcomes of decisions, relative to the distribution of possible alternatives.[9]

The second principle is that, like maps, data are *representational* objects, which means we can locate alternate representations to benchmark the value of a given data set. Indeed, data only has meaning in reference to the set of conceptual entities it describes, and there are (infinitely) other ways to represent such entities. We can liken the complete set of conceptual entities to a landscape or terrain from cartographic theory, be it a populace, transit space, or the gene-disease space from before. As with maps, a given data set is simply one 'representation' of this space. The unit of measure, level of observation, time interval, and granularity of the data may vary based on specific needs or decisions made by the data creators. Such choices may also reflect these creators' goals, incentives, and constraints. For example, a government data source is likely more detailed for easier-to-reach populations than harder-to-reach populations. Finally, the data represents the creator's understanding of a terrain as known at one point in time; this may change in the same way that emerging geographical data shaped the evolution of medieval maps to present-day world maps. The concept that data are not objective but are socially constructed representations is well known in the sociological and 'critical data studies' literature (Boyd & Crawford, 2012).

Hence, when relating decisions to a data representation, we are consciously mindful of other representations the data could have had. These representations represent the 'counterfactual' against which a particular data set's value will be measured. This counterfactual could be a blank map if no data is otherwise available to the decision-maker. It could be the decision-maker's private information, which could differ across decision-makers. It could be the best (or most likely used) alternate data set, which may have been compiled with meaningfully different choices. For instance, in our first mountaineering example, this would be the second-placed-winning map. Finally, the counterfactual could be constructed from the data set itself. For instance, if the data maps one part of the conceptual terrain with especially high fidelity, we could compare it with the standard used for the rest of the terrain. This corresponds to our second mountaineering example, where not all regions are equally well mapped. In each of these cases, this approach can provide the analyst with meaningful variation in the informational environment under which decisions are taken. The analyst should do well to keep in mind, however, that they can (at best) hope to infer the value generated by a *specific representation* of a terrain, which is not necessarily the objectively 'best' way to represent a terrain.

The third and final principle of the framework is that simply comparing decision-outcomes across different data representations is not sufficient, and we must instead rely on 'usable' (i.e., experimental) variation. Consider, for example, a district that frequently collects surveys about consumer behavior, sentiment, and

expectations throughout the district. Moreover, it is known that retail establishments regularly draw upon this data to inform their marketing decisions (e.g., where to open store locations or direct advertising). The analyst may wish to estimate the value of this sample data by comparing data-driven marketing decisions in urban and rural counties, knowing that sample data is more accurate in urban counties, resulting in a higher fidelity representation of consumers' spending habits. However, this analyst should beware: they will not be able to attribute the differences in marketing outcomes (e.g., product purchases) purely to the data because there are other salient differences between urban and rural regions that could be driving the difference. In other words, we must pay close attention to other differences between decision-making scenarios beyond the quality of the data; otherwise, our estimates of data value will also reflect these preexisting differences (the proverbial 'apples to oranges' comparison).

In an ideal scenario, the analyst could fully mitigate this concern by randomly assigning potential users into different informational environments: a treatment group (that receives access to one data representation) and a control group (that receives access to another or no data). In doing so, the analyst ensures that any other differences are randomized away in the process. To help make this point clear, consider the opposite process: a nonrandom assignment. For example, if the analyst assigns treatment instead based on each user's income level, then the treatment and control group will clearly differ in an important way (i.e., income), which may confound any estimate of the value of access to data. By using explicit random assignment, the analyst is statistically likely to produce two comparable groups. Indeed, to fix ideas, the probability of having only high-income users in one condition is exceedingly low if each individual has a 50/50 probability of being assigned to either condition.

In practice, however, conducting such experiments at scale can be prohibitively difficult and expensive, especially in our context where public data assets are often vast in size and impact. Hence, to overcome the "reflection problem" (Manski, 1993), the analyst can instead search for natural experiments that naturally simulate (or roughly mimic) a random assignment process without any manipulation on behalf of the analyst. How might such a process be naturally simulated? Angrist (1990) provides one famous example. In their study, they make use of the Vietnam-era draft lottery, which was used to determine the priority of draft-age men for induction into military service. Because the draft lottery was based on an individual person's birthday, the likelihood of a draft-age man serving was nearly random. In our case, if the analyst can find instances where, by pure coincidence (i.e., as if nature 'flipped a coin'), parts of a conceptual landscape are mapped differently, for instance, then they can be confident that the treatment and control groups will be *roughly* comparable, such that most of the difference in quality of decision-making can be attributed to differences in the representation. We will see more examples of what natural assignment processes in next section. Suffice to say, while discovering natural experiments is not always simple, it can be a powerful method when done right.

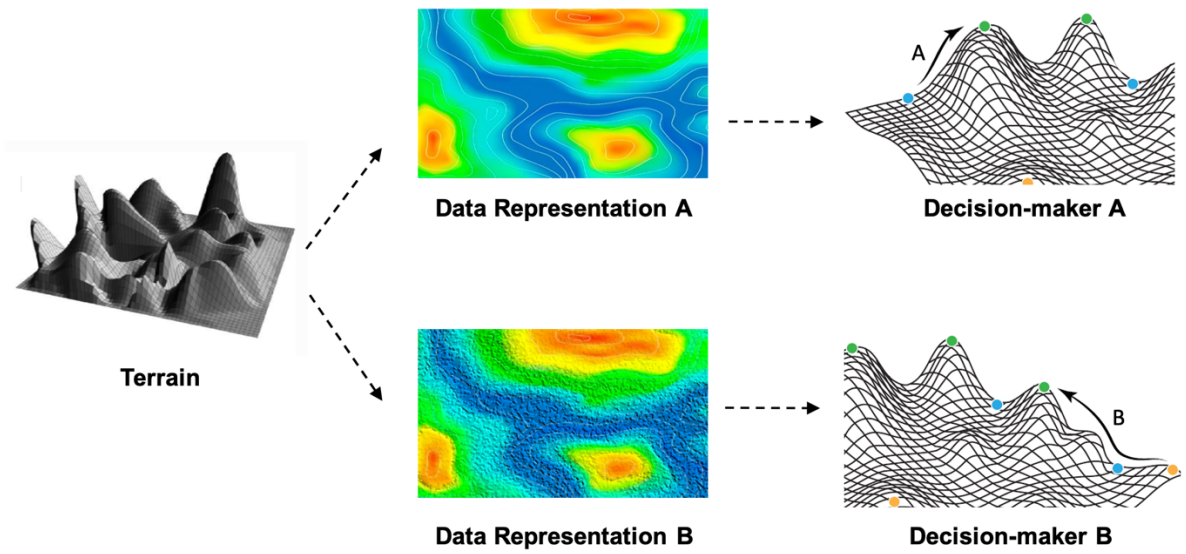

Figure 1. Stylistic Representation of Estimating Value Through Data Mapping

Figure 1 summarizes our framework. First, we must define the relevant terrain and the data-driven decisions that lie on top this terrain. This terrain is often the totality of conceptual entities a given data set hopes to describe. Second, we must discover alternate representations of the terrain; these alternate representations could reflect differences in coverage, access, cost, and so on. Third and finally, we must ensure that assignment to different representations is quasi-random. By comparing decisions made by agents on the landscape under the differing representations, we can reliably estimate the value of the data. If the natural experiment is convincing, then the only reason the agents should behave differently is because of differences in representation.

This approach addresses the key limitations introduced earlier. First, by focusing on data-driven decision-making instead of data usage, the analyst can bypass the myriad challenges associated with measuring data adoption. Indeed, we do not need to precisely understand who uses data in a given condition versus another. Rather, we must measure decision-making and outcomes in cases with alternate data representations. Second, this method captures the value that accrues to individuals who are influenced by the data, but do not use it directly, as long as we can measure *their* decision-making outcomes. Third, the diversity of applications is limited only by the diversity of outcomes we can capture concerning decision-making. Finally, the idea of counterfactuals is inherent to the framework, since the analyst has to be mindful about the variation they are using (i.e., variation subject to the quasi-random constraint), which means they will need to understand which representation the data's value is relative to.

3. Selected Examples of Quasi-Experiments for Public Data Valuation

This section provides four examples of the mapping framework when used in practice. Each case study uses quasi-experimental methods to infer data value based on concrete outcomes linked to some data terrain. Further, each case study exemplifies a distinct mode of data availability that is used to set up the natural experiment. The first example, drawn from the private sector, illustrates how public satellite data helps to uncover new market opportunities in the gold exploration industry. The second example describes how expanding access to public census-based data affects the rate and direction of research output in the economic and social sciences. The third example details how cost reductions in access to satellite data affect the quantity, quality, and diversity of environmental science. Finally, the fourth example estimates the value of online review listings using random data purchases by Yelp.[10]

3.1. The Private Impact of Public (Landsat) Data

A recent empirical study focusing on the $5 billion gold exploration industry (Nagaraj, 2021) illuminates the impact of public data on the private sector. For context, gold exploration is expensive and risky, and identifying a gold deposit may require years of exploration. Hence, mining companies have increasingly turned to public data from NASA and USGS's joint Landsat satellite mapping program for geological information that can help identify potential gold deposits. With this in mind, the study examines how Landsat data affects the rate of discovery of new gold deposits by exploration firms, as well as the share of discoveries made by market entrants (who now operate in a more level playing field) versus incumbents. In this example, the terrain is, quite literally, the surface of the Earth, and the key decision is the choice of which location to explore to discover new gold deposits.

Instead of surveying mining firms on their usage of Landsat imagery, or potentially simulating theoretically how discoveries might change if Landsat data were unavailable, the study takes advantage of idiosyncratic variation in Landsat coverage to set up a natural experiment. While NASA designed Landsat to have global coverage from launch, in practice, many regions did not receive Landsat imagery due to unexpected technical errors (e.g., data transmission errors) in data collection (see Figure 2a). Furthermore, even when images were collected, many locations received unusable data because the images were obscured by cloud cover, a notorious (and unpredictable) challenge for the remote sensing industry.

Per our conceptual framework, this provides us with variation in how well different parts of the terrain were represented: firms in locations where usable Landsat data was available ('treated regions') had a clearer map to detect gold mining potential, compared to locations where data was missing ('control regions'). The assignment process to each condition is already quasi-random because variation in data availability (i.e., transmission issues or cloud cover) should, in theory, have no direct relationship to gold mining potential, which means we can be confident we are isolating the long-term effects of operating in a location with strong

data coverage. However, we want to be fully confident that we are not ignoring any salient differences between the two kinds of regions and thereby making the dreaded apples-to-oranges comparison. Hence, to recover the causal effect of the satellite data, we instead compare whether the temporal increase in mining outcomes in the treated regions was *above and beyond* that experienced in the control regions, a commonly used method in the empirical social sciences.[11] Notably, the study focuses on the evolution of the mining *industry* within a geographic block (a region of 100 sq. miles that corresponds to one Landsat image) and tracks significant gold discoveries within each block. Given this level of analysis, the researcher does not need to rely on metrics of direct usage by individual mining firms. Further, it can capture spillovers to mining firms who reside in the location but do not use the data directly.

The findings are striking. Landsat imagery doubles the rate of gold discovery in a given region (off a small base) and increases the market share of junior firms in the industry from about 10% to 25%. Based on rough estimates of discovery value (derived from data on the size of discoveries), we estimate that the Landsat imaging program led to a gain of approximately $17 million for every mapped block over 15 years. For a country the size of the United States, this translates to additional gold reserves worth about $10 billion, which can be directly attributed to information provided by the Landsat program. This is an order of magnitude larger than the cost of operating and maintaining the program (in the range of a few hundred million dollars), and this is not even accounting for all other possible uses of the satellite imagery!

3.2. How Diffusion of Census Data Shapes Empirical Research

Access to public administrative data can affect the rate and direction of research in the social sciences and lead to more evidence-based policies. This was the finding of a recent systematic investigation (Nagaraj & Tranchero, 2023) that studied the downstream effects when researchers in the field of economics newly receive access to confidential census data.

Census microdata is the preeminent source of administrative data in the United States. It includes records of firm productivity, individual wages linked to firms, and individual-level demographic data (Abowd et al., 2004; Jarmin & Miranda, 2002). Yet, for privacy reasons, it is highly confidential, and access must be restricted. Traditionally, this meant that researchers intending to work with such data had to be physically present at the Census Bureau's headquarters in Maryland. However, starting in the 1990s, the bureau opened a network of secure data enclaves (termed Federal Statistical Research Data Centers, or FSRDCs) across the United States, enabling more researchers to access census data from new locations. In total, 30 FSRDCs were set up nationwide in a phased manner between 1994 and 2019 (see Figure 2b). Unlike in the Landsat example, where data access *within* a geography could change over time and space (e.g., as cloud cover changed), in this example, the data was consistently available in select geographies, but the *number* of geographies with access to this data changed over time. In other words, while the terrain (the national demography, economy, housing, etc.) was consistently well-mapped, not all researchers were given access to this map. Hence, from a methodological perspective, this study shows how to use variation in access restrictions to estimate data value.

In particular, the study explores how the opening of an FRSDC in a new city affects the research output of local researchers (our treatment group), compared with researchers for whom data access continues to be prohibitive (our control group). When embarking on a new research project, every researcher has to make the decision about which subject matter or phenomenon to focus on, and often she has to choose from multiple competing directions. This study's hypothesis is that more access to census data can help make this decision easier by showcasing many potential empirical analyses across a large swath of domains and providing early evidence of their merit. To the extent that this is true, access to census data can help researchers achieve their ultimate outcome of interest: namely, writing interesting papers that are accepted into high-impact journals, where they are cited extensively and are diffused into society to impact policies and human behavior. This is the value we are interested in estimating.

Of course, if the decision to open an FSRDC in a particular city were to depend on evolving trends in the research quality of the host academic institution (and other nearby institutions), then this research design would not be valid. There would be salient, preexisting differences between the two conditions that would confound estimates of data value. Fortunately, however, the decision to open an FSRDC in a given location was instead largely based on notions of equity and balance across geographies. Indeed, as one former FSRDC administrator explains, "Many institutions were interested in opening an RDC, but the NSF was interested in a kind of parity across the U.S. so that researchers in one part of the country had the same access as researchers in another part of the country did." (Interview T14, Nagaraj & Tranchero, 2023). This explains why, for instance, Stanford University only opened its FSRDC branch in 2010, given the long-standing presence of the nearby UC Berkeley FSRDC. This gave rise to quasi-random variation in who was designated with access to census data—a credible natural experiment to infer the value of census data to treated researchers.

We find that researchers newly exposed to confidential census data are far more likely to use (or build upon) the census data directly. Indeed, we estimate that the average researcher publishes 0.004 more papers in top journals using census data each year, which represents an increase of 131% from a baseline of 0.003 papers. This is not yet necessarily an outcome of interest. Rather, this shows that researchers are indeed factoring the data into their decision-making. However, we then *do* find evidence of multiple valuable outcomes that result from this: namely, researchers explore new questions that arise from the data, publish results in impactful papers in prestigious journals, and are also more likely to be cited in policy documents. This latter result is important since it highlights one potential channel for data to affect society writ large, influencing even government regulation and policies. Notably, our results are largely concentrated among empirical researchers (as opposed to theoretical researchers), which introduces a new form of inequality in the profession. For instance, empirical researchers published 0.035 more papers in top journals after accessing the data, an increase of 24% from a baseline of 0.145. For all these results, we treat all researchers residing in the same city as an FSRDC as being 'near it' (or, equivalently, 'exposed to' census data). However, we also test alternative measures of distance to an FSRDC (e.g., 1 mile, 2 miles, 5 miles, 10 miles). Our results become even stronger

the closer an economist is to an FSRDC and are strongest when researchers directly enjoy access on their campus. This suggests that our variation in data access is truly meaningful.

Using this design, we are also able to document significant spillovers to applied economists who do not use the data directly. For instance, when we remove direct FRSDC users from our analyses, our estimates of the effects of local access are 24%–39% smaller but remain large and significant. Further, we find that increasing awareness of past research using census data (or exposure to the findings of colleagues with access to FSRDCs) subsequently shapes the topics and questions that even non-exposed researchers decide to work on. Hence, not only does data access lead to new findings, but when knowledge of these findings diffuses, other researchers are inspired to produce more work. We would have missed this important second effect had the study relied on direct usage, even though it warrants inclusion in a holistic valuation of public data. In other words, direct data on the number of registered users of FSRDCs alone would not capture the full value of census data.

More generally, because academic insights are often a crucial driver of government policy, this study highlights how public data access (even under conditions as restrictive as FSRDCs) can positively impact research output and potentially affect economic policies downstream.

3.3. How Reduced Cost of Data Access Democratizes and Diversifies Scientific Research

The previous two examples show what happens when data access is temporally unavailable or closely restricted. Our third study (Nagaraj et al., 2020) investigates the effects when data access is *costly.* As in the first example, we once more leverage satellite imagery derived from the earlier-mentioned NASA Landsat program, which has the longest record of remote-sensing observations of the Earth. However, in this example, we examine the impact of reductions in costs and sharing restrictions for satellite imagery on environmental science research, a key beneficiary of remote sensing data. For almost a decade between 1985 and 1995, when the program was privatized, purchasing satellite images was prohibitively expensive, with a typical study costing at least $26,400. When the program was transferred back to the U.S. government, costs dropped significantly and data sharing became more relaxed. Using a sample of about 24,000 Landsat publications by over 34,000 authors matched to almost 3,000 unique study locations, we test how these cost changes affect the quality, quantity, and diversity of environmental science.

In this setting, the key decision that environmental scientists have to make is which region to choose to study place-based environmental phenomena, while (similar to the census data example) possible outcomes of interest include publishing success for the scientists themselves, as well as secondary broader effects for society (e.g., the diversity of research and how well regions are represented in the field). Since this decision is (naturally) informed by which regions have high-quality remote-sensing data available, the study once more exploits the fact that Landsat coverage was not uniform to set up another natural experiment. In particular, the

technical errors and cloud cover resulted in some regions already receiving consistently poor satellite image quality *even before commercialization*, which would have made it difficult to study these select regions. While regions with usable imagery were, in theory, more agreeable to study, the exorbitant costs during the commercial period would have presented its own difficulties. Yet, once commercialization concluded, we contend that only the latter regions would have been the beneficiaries of the reduced data costs. We therefore designate these regions with a greater amount of usable imagery as our treatment group, and regions with a lesser amount as our control group.

To operationalize this, we record the number of high-quality images in each geographic block in the year 1985 (the first year of commercialization) and split the sample into blocks with above-median coverage (which we consider the treatment group) and blocks with below-median coverage (which serves as the control). While it may seem redundant to have had multiple images for the same block, in reality this was essential for authors to study dynamic changes in environments, such as urbanization and deforestation. Once more, to be mindful of any salient differences between the two kinds of regions, we estimate whether treated regions experienced a greater increase in post-commercialization when compared with control regions. Altogether, this is a more credible research design than, for instance, simply comparing the total quantity (and quality) of scientific publications during and after the commercial period, since these may have evolved independently of the cost changes.

We find that reduced costs of satellite imagery substantially affect the quantity and quality of Landsat-enabled science. In particular, we estimate that the number of yearly published research articles pertaining to a given geographic block increases by a factor of 3, while the probability of any publication increases by about 50%. The number of highly cited publications increases by a factor of 6. Because an open data regime is likely to benefit scholars endowed with less financial resources, one might also expect much of these gains to originate from scientists in lower-ranked institutions and the developing world. We do indeed find this to be the case. There is a material difference between the increase in publications we observe from authors in lower-ranked institutions (Top 50–200 in the Quacquarelli Symonds World Rankings) as compared with scholars in the top-ranked institutions (Top 50). There is a similar differential impact between scholars from low-income countries and scholars in the United States and other high-income countries (based on World Bank income data). This democratization, in turn, increases the geographic and topical diversity of Landsat-enabled research (see Figure 2c, which highlights all new study locations postcommercialization). Furthermore, scientists are more likely to focus on previously understudied regions close to their home location or introduce otherwise novel research topics. These findings suggest that policies that improve access to valuable scientific data promote scientific progress, reduce inequality among scientists, and increase the diversity of scientific research.

Methodologically, this study advances the use of cost channels as a key mechanism to estimate the effects of public data. More generally, this work highlights how the value of public data can be distributed across

researchers of different status. Should the policymaker place more weight on stimulating a diverse set of follow-on applications, they may consider attributing extra value to the data.

3.4. How Bulk Data Drops of Online Listings Affect Business Performance

Our fourth and final study shows how an analyst can leverage improvements in data coverage and data purchases to estimate the value of data. In this study, Luca et al. (2022) investigate how online listing platforms drive business outcomes, with a particular focus on Yelp. Yelp is a major online review platform with the potential to facilitate the growth of small businesses, especially restaurants and bars, by enabling them to develop an online reputation. However, not all establishments maintain a presence on Yelp, even though online pages are free to set up and easy to do so. This may be because some business owners simply lack awareness of the value of an online presence, or more severely, they may fear the potential backlash from negative reviews.

The study focuses on Texas as a case study, merging state tax data on restaurants and bars with proprietary data from Yelp on local businesses, their revenues, and the dates of listings. Using this data, we first simply look at what happens when any establishment in the state of Texas is added to Yelp, and find that, on aggregate, this is correlated with an increase in sales by over 5%. Note that this design is not (yet) a natural experiment: while it is tempting to think we have a natural treatment group (i.e., establishments added to Yelp) and a control group (establishments that remain off Yelp), these two groups likely differ in meaningful ways. In particular, we might worry that high-performing establishments that anticipate future success may be overhauling their marketing strategy, which includes adding themselves to Yelp. This would mean that the control group does not provide a suitable counterfactual. Nevertheless, the study does control for several important factors, including some contemporaneous business changes (e.g., ownership changes and advertising spending), and the locations establishments operate in.

Recall the earlier discussion regarding the gap between how a terrain is represented and its ground truth: in this context, the iterative addition of listings to Yelp's database creates a map with better coverage (see Figure 2d for how the coverage of establishments on the platform changed over time). In one important instance of this, Yelp worked with a local business data aggregator in Texas to add over a thousand local restaurants and bars in the space of two days (a 'bulk' addition). This now does produce a credible natural experiment: these 'bulk' establishments did not voluntarily list themselves (creating possible biases), yet they still happened to be added to Yelp simply because they were already in some local aggregator's database. They constitute the treatment group, while the businesses that remained off the platform now represent a suitable control group. From this natural experiment, we find that the impact of a bulk addition results in a 10% positive and significant increase in revenues, which is even greater than the aggregate effect of 5% for the state of Texas. For comparison, the average quarterly revenue for these establishments is approximately $60,000, so the effect size is highly economically significant.

The study shows how data purchases can lead to a more complete map while simultaneously providing usable variation that we can use to estimate the value of data for decision-making. Studies of this nature can also help business owners assess the value of online listings, as well as guide policymakers who are trying to encourage small businesses to maintain a digital presence, like the Organisation for Economic Co-operation and Development's recent Small and Medium Enterprises Global Initiative.[12] In many other contexts, data is constantly being improved, and coverage keeps expanding over time. These events can act as useful natural experiments to study the impact of better coverage on affected units.

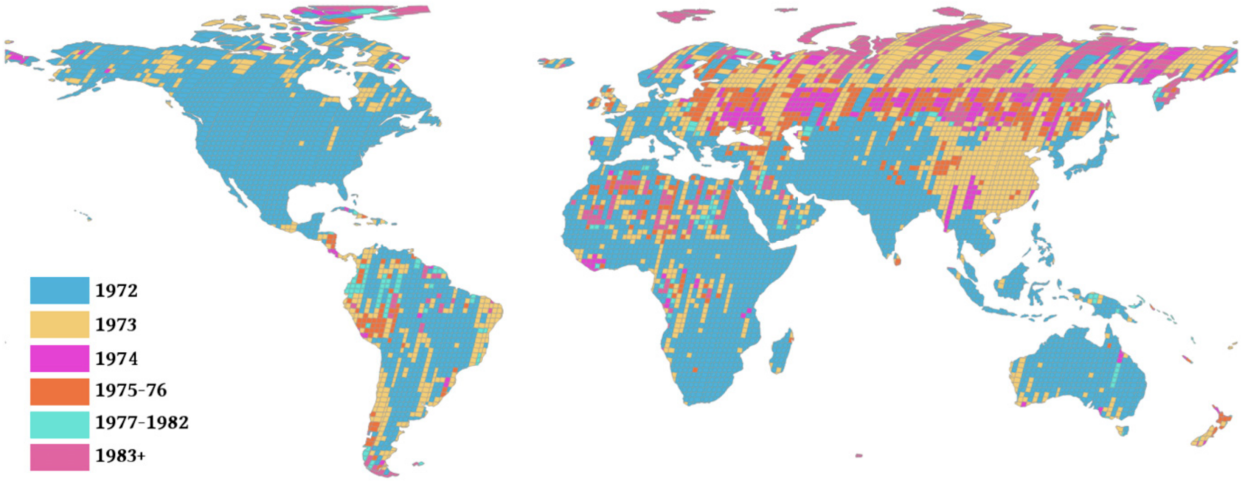

Figure 2a. Example of Patchy Coverage of Early Landsat Satellite Imagery

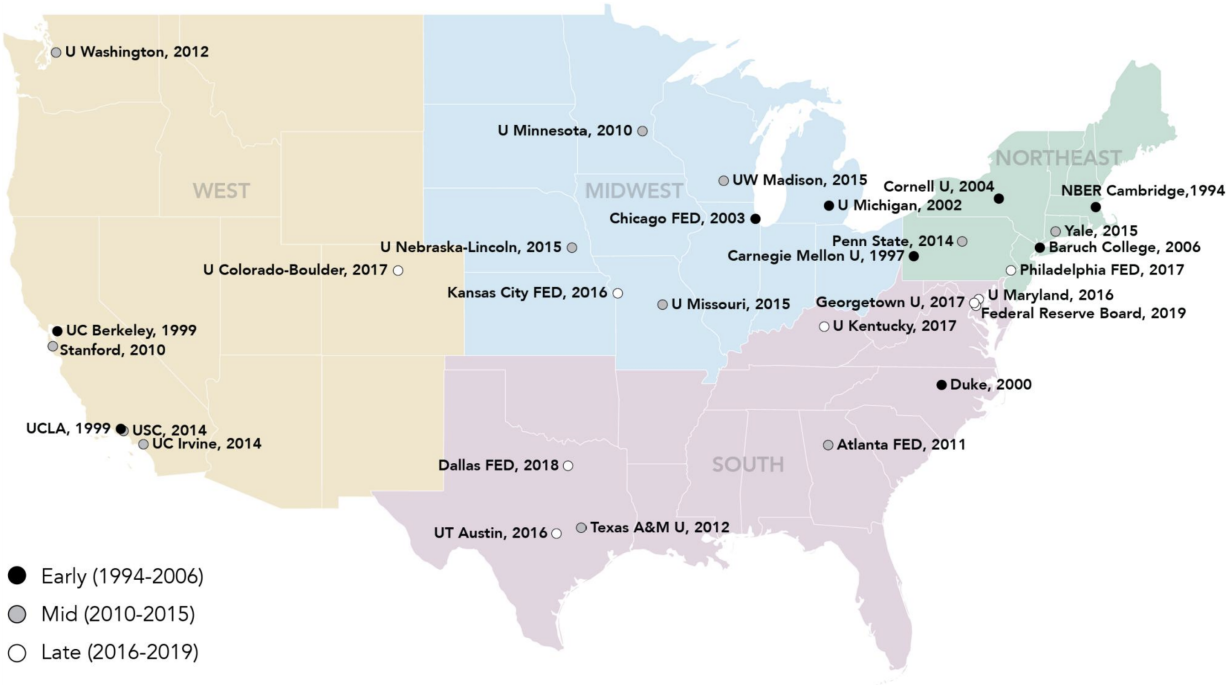

Figure 2b. Gradual Diffusion of Federal Statistical Research Data Centers in the United States

Figure 2c. The Geography of Science Using Landsat Imagery Before and After Commercialization

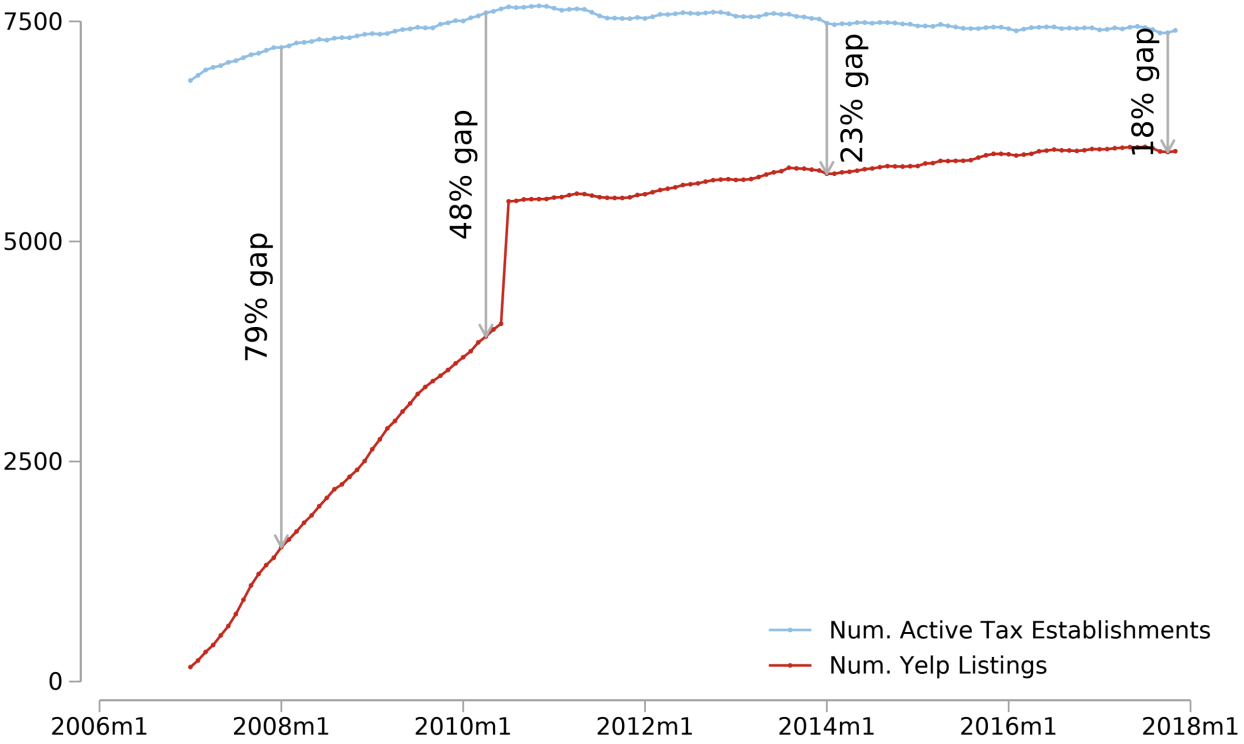

Figure 2d. Variation in Yelp Coverage Over Time

Table 1. Overview of studies for the quasi-experimental methods for estimation of data value.

Question	Quasi-Experimental Variation	Outcome

Satellite data in the gold exploration industry	Patchy coverage due to technical difficulties and cloud cover during satellite data collection	Satellite data doubles discoveries and encourages the entry of new gold exploration firms
Census administrative data and economics research	Changing access at the university level given the need to access data via physical enclaves	Access to administrative data leads to more top publications by empirical researchers, and even among those who do not use the data directly
Satellite data cost and environmental science	Changing cost in access to data given changing government regulation	Lowering the cost of data access leads to more diversity in the set of regions and scientists that benefit
Restaurant listings and revenue on Yelp	Increased coverage of restaurant listings owing to a data purchase by Yelp	Improved coverage increases quarterly revenue by 5–10%

Table 1 summarizes the four studies mentioned above, as well as the key sources of variation. In each of these settings, it is important to remember that the data is a map of the phenomena it describes and that we use quasi-experimental variation in assignment to the treatment condition to estimate its value. Our hope is that the analyst can draw from these examples to evaluate data value in other contexts using this mapping lens.

4. Comparisons to Alternative Approaches to Data Valuation

The quasi-experimental approach is not the only technique researchers can use to estimate the value of public data. Indeed, they may also be able to use surveys, simulations, or experiments. We briefly review these other empirical techniques below.

4.1. Surveys

Arguably the most straightforward approach to assessing data value is to survey users currently using the data. With this method, users are periodically asked about the specific applications of the given data set in their work, which subset of the overall data they find useful, why this might be the case, and what other data they would like to be made available. These surveys can provide a comprehensive review of all possible applications of a particular data set, since they are not limited to any one specific application. For instance, Loomis et al. (2015) survey users across the government, private sector, and academia to estimate the monetary value of Landsat images. As such, surveys can help to capture unexpected uses of the data not foreseen by the analyst.

Nevertheless, surveys can also underestimate the data value that is created for end users. This is because it is difficult for survey respondents to imagine counterfactual outcomes (i.e., what they might have done had the data not been available) and the magnitude of outcomes under these counterfactuals. In addition, it can be challenging to evaluate the cost savings of public data (e.g., in comparison with more expensive commercial

data) using survey data. This comparison assumes that decision-makers would use the same amount or type of data had it been more expensive. Instead, the amount of information used is likely to vary with the cost, yet the researcher cannot observe this relationship using survey data alone.

4.2. Simulations

While surveys can be excellent tools to capture potential data uses, ultimately, they cannot fully capture how data affects decisions, nor how these decisions might change when access to data is unavailable. One alternative method that potentially addresses these challenges is to set up a theoretical model that explicitly specifies how data is incorporated into the decision-making in a specific decision context. The analyst can then simulate different amounts of data, or data of varying accuracy and frequency, to predict the possible decisions and the corresponding utilities of these decisions. The Democratizing Data Initiative, as described by Emecz et al. (2024), exemplifies this approach. It analyzes the impact of data assets from U.S. federal agencies, employing machine learning algorithms and subject matter expert validation to assess data utilization by the research community. Another example of this approach is seen in Baquet et al. (1976), which studies the importance of seasonal winter weather forecasts for farming. To derive their estimates, the study authors set up a model combining a monetary payoff table (based on expected yields, production costs, and crop values) with a forecast distribution (derived from forecast information and historical data) to simulate probabilistic values. This approach has gained increasing recognition in recent years among practitioners. For instance, in a recent study, Lauer et al. (2021) estimate the societal benefits of improving the NOAA's next generation geostationary satellite program (GeoXO). In the study, the authors note the potential of "observing system simulation experiments (OSSEs)," which feed simulated data into numerical weather models to produce operational forecasts. They also provide a micro-economic framework to model the use of these forecasts by decision-makers in weather contexts.

The advantage of this approach is that it carefully specifies how data affects value and enables the analyst to infer the value of such data through sound economic logic. Furthermore, this approach also makes it possible to simulate the value of data even before it is generated and subsequently use the resulting calculations to invest in data-generating technologies or projects. Nevertheless, this approach is also not without its limits. First, it tends to result in a narrow focus on a few applications (i.e., ones the analyst can model theoretically) and, almost by definition, it tends to ignore unexpected uses. Second, because the estimates rely on simulations, the realized outcomes of data-driven decisions may deviate from predicted outcomes should the participants not act per the assumptions in the specified theoretical model. For example, many of these theoretical models tend to assume profit maximization. To the extent that participants exhibit behavioral biases, there may be errors in the estimates of the value of information. Hence, simulation-based approaches are most valuable when the underlying theoretical model can be fully specified, validated, and considered trustworthy. While less common, simulation-based approaches can also be used in combination with surveys to estimate the value of information.

4.3. Experimental Approaches

The final method (and the one closest to our approach) is to implement explicit field or lab experiments to estimate data value. This approach can take advantage of controlled conditions, using random assignment to designate who is given access to specific data and who is not. For example, in a decision-making experiment, a group of participants may receive a comprehensive data set to aid their decisions, while another group may have to rely only on intuition or less information. If the experiment is designed well, the comparison of outcomes between the two groups would help to illuminate the value of the data. For example, in a recent working paper, Pakhtigian et al. (2024) explore how households respond to personalized cholera risk predictions to modify their water usage habits. The investigation draws upon a 9-month field experiment that involves a smartphone application that provides individualized monthly cholera risk forecasts linked to the user's residential location. The authors find that households given access to the application feel more prepared to tackle potential environmental and health hazards.

In a lab experiment, Hoelzemann et al. (2022) explore the potential *adverse* effects of data access, and in particular the 'streetlight effect.' This is when data uncovers a somewhat attractive project (but not the most attractive project), which can preclude exploration in search of the more attractive projects. Theoretically, this could reduce social welfare even relative to cases when no data is provided at all. This effect is named after the proverbial case of a drunkard searching for a lost car key under the streetlight since it is the only illuminated place. To test this phenomenon empirically, the authors employ an online lab experiment that asks participants to search for hidden gems whose monetary values are either low, medium, or high. In one set of conditions, the location of the medium-value gem is revealed, while in another, the participants have to make decisions without any data. They find that participants earn approximately 12% less under the "streetlight" condition where data on the medium outcome is provided. While knowing which option harbors the medium outcome improves payoffs in the short run (i.e., in earlier rounds), it reduces the likelihood that the high-value gem will be discovered, lowering overall payoffs in the long-run. This study shows that the value of data is not necessarily unidirectional. Rather, it can either stimulate or crowd out social learning, which means an analyst could find that a given data set actually has a *negative* value. The study also provides a useful experimental framework to assess the value of data in the lab.

5. Discussion

Public data is a critical resource for many scientific and commercial activities. Increasingly aware of this potential (and the role of public data as a public good), government institutions have become more thoughtful about when to (and how to) collect, store, analyze, and disseminate public data. Indeed, going forward, amid growing debate surrounding data privacy, agencies will need to balance the utility of public data provision against the risk of incurring privacy losses (using aptly called 'risk-utility' frameworks). While the concept of privacy losses can be relatively abstract, and prior work has already established credible approaches to quantify these, our methodology provides a new approach to *concrete*ly measure utility and assign a dollar measure to

critical public data assets. In general, a more comprehensive and accurate understanding of data value will help the federal government navigate the new terrain of evidence-based decision-making.

However, our understanding of (and attempts to estimate) data value has so far been hindered by several challenges, including the difficulties associated with determining precise usage of data, as well a general incompleteness of data over time and space. To overcome these challenges, we introduce a novel approach in this article for assessing data value. The heart of this approach is centered around conceptualizing data as a map —a metaphor that clarifies the value of data in decision-making contexts. Using four case studies, we demonstrate how the value of data can be estimated using natural experiments that exploit differences in data representations, be it through coverage, accessibility, cost, or enhancements through data purchases. This perspective can aid users in comprehending the terrain of decision-making, as well as equip them to fill in gaps where necessary. Unlike traditional methods for data valuation, such as surveys or simulations, our approach can help provide a more comprehensive and rigorous examination of data's real-world utility and influence. However, we caution that while this methodology can yield a richer understanding, it might require more time for execution, and assigning precise dollar values to data may involve additional assumptions (e.g., the dollar value of an academic citation).

How might a data publisher begin to use our framework? Consider an institution with two flagship data sets (data A and data B) that is deciding which to prioritize for a series of improvements. While, theoretically, they should invest in the data improvement that delivers the highest net present value (NPV), this is perhaps infeasible to estimate, and a more natural starting point would be to simply prioritize the data set with higher data value. Of course, as we have learned, this is not a trivial problem in and of itself. If lucky, the publisher may be able to locate one natural experiment whose alternative representations already correspond to the two data sets of interest. Then, the publisher will be able to infer the relative value of each data set in one step. More commonly, however, they will need to design two separate natural experiments, potentially leveraging independent sources of variation (e.g., access restrictions to estimate the value of one data set, and cost restrictions to estimate the value of the other).

In another scenario, consider a government agency that is deciding whether or not to publish a critical data asset. They have already used a differential privacy approach to quantify the privacy losses associated with the public release of the data set, however, they still need to understand the corresponding 'utility.' If the institution has already granted access to a (restricted) group of decision-makers, then they may be able to estimate the data value from the outcomes of this group's decisions. Often, it is a good idea to focus on several (large) applications, locate a natural experiment for each, and finally calculate the composite value. If there is no history of data usage outside of the agency, then this agency may need to look for similar data sets in other contexts, whose values have already been estimated using natural experiments. The agency could then compare the quantified privacy losses with the distribution of data values. We leave other possible implementations of our framework for future research.

In conclusion, we believe that, despite its vast potential, public data remains largely untapped and underfunded. In our experience, the failure to recognize the true value of this data is partly attributable to the inability to account for the spillovers and other secondary effects of data. Therefore, we must shift our focus toward the totality of benefits provided by public data and make use of methods that can capture this full value. The principles outlined in this article present a compelling pathway toward this goal. With greater appreciation for public data's value, we can foster more informed decisions about data generation, dissemination, and utilization, thereby maximizing the role of data as a public good.

Acknowledgments

This article benefited from guidance and encouragement from Julia Lane. Cecil-Francis Brenninkmeijer, Nilo Mitra, and Yanqi Cheng provided excellent research assistance. All errors are my own.

Disclosure Statement

Abhishek Nagaraj has no financial or non-financial disclosures to share for this article.

References

Abowd, J. M., Haltiwanger, J., & Lane, J. (2004). Integrated longitudinal employer-employee data for the United States. *American Economic Review*, *94*(2), 224–229. https://doi.org/10.1257/0002828041301812

Angrist, J. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security administrative records. *American Economic Review*, *80*(3), 313–336.

Baquet, A., Halter, A., & Conklin, F. (1976). The value of frost forecasting: A Bayesian appraisal. *American Journal of Agricultural Economics*, *58*(3), 511–520. https://doi.org/10.2307/1239268

Bloom, N., Schankerman, M., & Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, *81*(4), 1343–1393. https://doi.org/10.3982/ECTA9466

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brynjolfsson, E., Eggers, F., & Gannamaneni, A. (2018). *Using Massive Online Choice Experiments to Measure Changes in Well-Being.* https://doi.org/10.3386/w24514

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, *43*(2), 245–257. https://doi.org/10.1177/001979399004300205

Card, D., & Krueger, A. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *The American Economic Review*, *84*(4), 772–793. https://doi.org/10.3386/w4509

Casaburi, L., & Troiano, U. (2016). Ghost-house busters: The electoral response to a large anti-tax evasion program. *The Quarterly Journal of Economics*, *131*(1), 273–314. https://doi.org/10.1093/qje/qjv041

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Eds.), *TCC 2006: Theory of cryptography*. Lecture Notes in Computer Science (Vol. 3876, pp. 265–284). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11681878_14

Emecz, A., Mitschang, A., Zdawcyk, C., Dahan, M., Baas, J., & Lemson, G. (2024). Turning visions into reality: Lessons learned from building a search and discovery platform. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.d8a3742f

Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2019). https://www.congress.gov/bill/115th-congress/house-bill/4174

Hausen, R., & Azarbonyad, H. (2024). Discovering data sets through machine learning: An ensemble approach to uncovering the prevalence of government-funded data sets. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.18df5545

Hoelzemann, J., Nagaraj, A., Manso, G., & Tranchero, M. (2022). *The streetlight effect in data-driven exploration. [Working Paper]*. University of California, Berkeley. https://static1.squarespace.com/static/6232322165ae714579213636/t/6359a587227cca7a131509b2/1666819471534/the-streetlight-effect-in-data-driven-exploration.pdf

Imbens, G., Rubin., D., & Sacerdote, B. (1999). *Estimating the effect of unearned income on labor supply, earnings, savings, and consumption: Evidence from a survey of lottery players*. NBER Working Paper No. 7001. National Bureau of Economic Research. https://doi.org/10.3386/w7001

Jarmin, R. S., & Miranda, J. (2002). *The longitudinal business database*. SSRN. http://dx.doi.org/10.2139/ssrn.2128793

Korzybski, A. (1993). *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. International Non-Aristotelian Library Publishing Company.

Laxminarayan, R., & Macauley, M. (2012). *The value of information: Methodological frontiers and new applications in environment and health.* Springer Dordrecht.

Lauer, C., Conran, J., & Adkins, J. (2021). Estimating the societal benefits of satellite instruments: Application to a break-even analysis of the GeoXO Hyperspectral IR Sounder. *Frontiers in Environmental Science*,

9. https://doi.org/10.3389/fenvs.2021.749044

Luca, M., Nagaraj, A., & Subramani, G. (2022). *Getting on the map: The impact of online listings on business performance*. NBER Working Paper No. 30810. National Bureau of Economic Research. https://doi.org/10.3386/w30810

Loomis, K., Koontz, S., Miller, H., & Richardson, L. (2015). Valuing geospatial information: Using the contingent valuation method to estimate the economic benefits of Landsat satellite imagery. *Photogrammetric Engineering & Remote Sensing*, *81*(8), 677–656. https://doi.org/10.14358/pers.81.8.647

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), 531–542. https://doi.org/10.2307/2298123

Mashima, D., Kobourov, S., & Hu, Y. (2012). Visualising dynamic data with maps. *Institute of Electrical and Electronics Engineers*, *18*(9), 1424–1437. https://doi.org/10.1109/TVCG.2011.288

Nagaraj, A. (2021). The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry. *Management Science*, *68*(1), 564–582. https://doi.org/10.1287/mnsc.2020.3878

Nagaraj, A., Shears, E., & De Vaan, M. (2020). Improving data access democratizes and diversifies science. *Proceedings of the National Academy of Sciences*, *117*(38), 23490–23498. https://doi.org/10.1073/pnas.2001682117

Nagaraj, A., & Stern, S. (2020). The economics of maps. *Journal of Economic Perspectives*, *34*(1), 196–221. https://doi.org/10.1257/jep.34.1.196

Nagaraj, A., & Tranchero, M. (2023). *How does data access shape science? Evidence from the impact of U.S. Census's research data centers on economics research*. NBER Working Paper No. 31372. National Bureau of Economic Research. https://doi.org/10.3386/w31372

Pakhtigian, E., Aziz, Boyle, K., J., Akanda, A., S., & Hanifi, S.M.A. (2024). Early warning systems, mobile technology, and cholera aversion: Evidence from Rural Bangladesh. *Journal of Environmental Economics and Management*, *125*, Article 102966. https://doi.org/10.1016/j.jeem.2024.102966

Potok, N. (2024). Data usage information and connecting with data users: U.S. mandates and guidance for government agency evidence building. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.652877ca

Simon, B., Crowley, C., & Franco, F. (2022). The costs and costs avoided from wildfire fire management—A conceptual framework for a value of information analysis. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.804958

Sostek, K., Russell, D., Goyal, N., Alrashed, T., Dugall, S., & Noy, N. (2024). Discovering datasets on the web scale: Challenges and recommendations for Google Dataset Search. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.4c3e11ca

The Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel 2021. NobelPrize.org. https://www.nobelprize.org/prizes/economic-sciences/2021/summary/

Tranchero M. (2023). *Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation* [Working Paper]. University of California, Berkeley. https://www.matteotranchero.com/pdf/Matteo_Tranchero_JMP_latest.pdf

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Footnotes

1. We also include other sources of (freely available) data provided either by firms (e.g., Google Trends or Maps data) or nonprofits (such as the Wikidata project). ↩
2. See, for instance, Dwork et al. (2006). ↩
3. In theory, we could extend a similar logic to design randomized control trials (RCTs) to test the value of data. However, such experiments are often narrow in scope and may be infeasible for many important, large-scale data sources. They may also present ethical concerns, particularly when exposure to treatment can be life-altering (e.g., access to early warning hurricane forecasts). We return to RCTs in more detail in section 4.3. ↩
4. In response, some institutions, like the National Oceanic and Atmospheric Agency (NOAA), now require the user to force a service delivery request, which allows the agency to learn the user profile, as well as capture novel applications, such as retailers using climate forecasts to stock winter gear. ↩
5. See, for example, Brynjolfsson et al. (2018) ↩
6. See https://www.rff.org/valuables/ for more examples. ↩
7. Often, the term ‘map’ is used more generally to refer to any abstract representation of physical or intangible objects in space. This may be broader than geographic features (e.g., lakes and roads) and also

include demographic, sociological, and economic characteristics that are arrayed within a space. See, for example, Mashima et al. (2012). However, we believe that, for our purposes, the narrow, cartographic definition of maps is most useful for the reader to understand the framework. ↩

8. Beyond their role in individual decision-making, maps also play a role in commercial and political decision-making – in industrial organization via the location of transportation hubs, factories, and customers; in public finance via topographical, census, tax, insurance, and weather maps; in political economy (via policies on gerrymandering and property rights); and in housing and financial markets via land survey, redlining, and flood insurance maps. ↩

9. This should not imply that individuals always make the best use of data in their decision-making processes. Indeed, a wide literature has explored the role of cognitive biases in human decision-making—see, for instance, Tversky & Kahneman (1974)—and many of these biases can interact in important ways with data (e.g., ‘anchoring' to a data-driven estimate). Nevertheless, while humans may not maximize the value of data, to the extent that there is any value, we argue that it must be derived from the outcomes of decisions taken with this data. ↩

10. In each setting, we highlight the value of improved access in terms of contextual outcomes (e.g., more gold discoveries, more research papers, more revenues, etc.). We need additional assumptions germane to the situation or sector of interest to translate each outcome into dollar terms, but this is generally doable. ↩

11. In more technical terms, we estimate a difference-in-differences framework using two-way fixed effects. We do so in all four examples; however, for ease of understanding, we sometimes simply refer to this as comparing treatment and control groups. ↩

12. See https://www.oecd.org/digital/sme/ ↩

References

- Abowd, J. M., Haltiwanger, J., & Lane, J. (2004). Integrated longitudinal employer-employee data for the United States. *American Economic Review*, *94*(2), 224–229. https://doi.org/10.1257/0002828041301812 ↩
- Angrist, J. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security administrative records. *American Economic Review*, *80*(3), 313–336. ↩
- Baquet, A., Halter, A., & Conklin, F. (1976). The value of frost forecasting: A Bayesian appraisal. *American Journal of Agricultural Economics*, *58*(3), 511–520. https://doi.org/10.2307/1239268 ↩

- Bloom, N., Schankerman, M., & Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, *81*(4), 1343–1393. https://doi.org/10.3982/ECTA9466 ↩
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878 ↩
- Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, *43*(2), 245–257. https://doi.org/10.1177/001979399004300205 ↩
- Card, D., & Krueger, A. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *The American Economic Review*, *84*(4), 772–793. https://doi.org/10.3386/w4509 ↩
- Casaburi, L., & Troiano, U. (2016). Ghost-house busters: The electoral response to a large anti-tax evasion program. *The Quarterly Journal of Economics*, *131*(1), 273–314. https://doi.org/10.1093/qje/qjv041 ↩
- Emecz, A., Mitschang, A., Zdawcyk, C., Dahan, M., Baas, J., & Lemson, G. (2024). Turning visions into reality: Lessons learned from building a search and discovery platform. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.d8a3742f ↩
- Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2019). https://www.congress.gov/bill/115th-congress/house-bill/4174 ↩
- Hausen, R., & Azarbonyad, H. (2024). Discovering data sets through machine learning: An ensemble approach to uncovering the prevalence of government-funded data sets. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.18df5545 ↩
- Hoelzemann, J., Nagaraj, A., Manso, G., & Tranchero, M. (2022). *The streetlight effect in data-driven exploration. [Working Paper]*. University of California, Berkeley. https://static1.squarespace.com/static/6232322165ae714579213636/t/6359a587227cca7a131509b2/1666819471534/the-streetlight-effect-in-data-driven-exploration.pdf

↩

- Imbens, G., Rubin., D., & Sacerdote, B. (1999). *Estimating the effect of unearned income on labor supply, earnings, savings, and consumption: Evidence from a survey of lottery players*. NBER Working Paper No. 7001. National Bureau of Economic Research. https://doi.org/10.3386/w7001

↩

- Jarmin, R. S., & Miranda, J. (2002). *The longitudinal business database*. SSRN. http://dx.doi.org/10.2139/ssrn.2128793

↩

- Korzybski, A. (1993). *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. International Non-Aristotelian Library Publishing Company.

↩

- Lauer, C., Conran, J., & Adkins, J. (2021). Estimating the societal benefits of satellite instruments: Application to a break-even analysis of the GeoXO Hyperspectral IR Sounder. *Frontiers in Environmental Science*, *9*. https://doi.org/10.3389/fenvs.2021.749044

↩

- Laxminarayan, R., & Macauley, M. (2012). *The value of information: Methodological frontiers and new applications in environment and health.* Springer Dordrecht.

↩

- Loomis, K., Koontz, S., Miller, H., & Richardson, L. (2015). Valuing geospatial information: Using the contingent valuation method to estimate the economic benefits of Landsat satellite imagery. *Photogrammetric Engineering & Remote Sensing*, *81*(8), 677–656. https://doi.org/10.14358/pers.81.8.647

↩

- Luca, M., Nagaraj, A., & Subramani, G. (2022). *Getting on the map: The impact of online listings on business performance*. NBER Working Paper No. 30810. National Bureau of Economic Research. https://doi.org/10.3386/w30810

↩

- Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), 531–542. https://doi.org/10.2307/2298123

↩

- Nagaraj, A. (2021). The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry. *Management Science*, *68*(1), 564–582. https://doi.org/10.1287/mnsc.2020.3878

↩

- Nagaraj, A., & Stern, S. (2020). The economics of maps. *Journal of Economic Perspectives*, *34*(1), 196–221. https://doi.org/10.1257/jep.34.1.196

  ↩

- Nagaraj, A., & Tranchero, M. (2023). *How does data access shape science? Evidence from the impact of U.S. Census's research data centers on economics research*. NBER Working Paper No. 31372. National Bureau of Economic Research. https://doi.org/10.3386/w31372

  ↩

- Nagaraj, A., Shears, E., & De Vaan, M. (2020). Improving data access democratizes and diversifies science. *Proceedings of the National Academy of Sciences*, *117*(38), 23490–23498. https://doi.org/10.1073/pnas.2001682117

  ↩

- Pakhtigian, E., Aziz, Boyle, K., J., Akanda, A., S., & Hanifi, S.M.A. (2024). Early warning systems, mobile technology, and cholera aversion: Evidence from Rural Bangladesh. *Journal of Environmental Economics and Management*, *125*, Article 102966. https://doi.org/10.1016/j.jeem.2024.102966

  ↩

- Potok, N. (2024). Data usage information and connecting with data users: U.S. mandates and guidance for government agency evidence building. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.652877ca

  ↩

- Simon, B., Crowley, C., & Franco, F. (2022). The costs and costs avoided from wildfire fire management—A conceptual framework for a value of information analysis. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.804958

  ↩

- Sostek, K., Russell, D., Goyal, N., Alrashed, T., Dugall, S., & Noy, N. (2024). Discovering datasets on the web scale: Challenges and recommendations for Google Dataset Search. *Harvard Data Science Review*, (Special Issue 4). https://doi.org/10.1162/99608f92.4c3e11ca

  ↩

- *The Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel 2021*. NobelPrize.org. https://www.nobelprize.org/prizes/economic-sciences/2021/summary/

  ↩

- Tranchero M. (2023). *Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation [Working Paper]*. University of California, Berkeley.

↩