

ARTICLE

What is the value of data? A review of empirical methods

Diane Coyle  | Annabel Manley

Bennett Institute for Public Policy,
University of Cambridge, Cambridge, UK

Correspondence

Diane Coyle, Bennett Institute for Public
Policy, University of Cambridge,
Cambridge, UK.
Email: dc700@cam.ac.uk

Funding information

Omidyar Network

Abstract

With the growing use of digital technologies, data have become core to many organizations' decisions, with its value widely acknowledged across public and private sectors. Yet few comprehensive empirical approaches to establishing the value of data exist, and there is no consensus about which methods should be applied to specific data types or purposes. This paper examines a range of data valuation methodologies proposed in the existing literature. We propose a typology linking methods to different data types and purposes.

KEYWORDS

assets, data types, value of data

JEL CLASSIFICATION

C80, O30

1 | INTRODUCTION

The economy has been transformed by data in recent years. Data-driven firms made up seven of the global top 10 firms by stock market capitalization in 2021; and across the Organization for Economic Co-operation and Development (OECD) economies there has been a growing gap in terms of productivity and profitability between firms that use data intensively and the rest (e.g., Brynjolfsson et al., 2021; Bajgar et al., 2021; Coyle et al., 2022). Examples of valuable uses vary widely, including for example data analytics for marketing or logistics, improved process automation, fraud detection in finance, and increasingly applications of machine learning or artificial

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Journal of Economic Surveys* published by John Wiley & Sons Ltd.

intelligence (Veldkamp & Chung, 2022). The widespread availability of data and analytics has also begun to extend into the public sector and policymaking; for example, “following the science”—implying intense use of data—becoming a tagline for the handling of the COVID-19 pandemic in the UK and elsewhere.

It is therefore obvious that data have value in an economically meaningful sense. The sources of its value and characteristics of data as an economic asset are discussed at length in Coyle et al. (2020a). They observed that there is potential social welfare value to the economy as a whole, and not just private value to the organizations that control specific data sets, from having the ability to use data. This appreciation is increasingly reflected in many policy statements of data strategy and the broader international policy debate about the governance of data (e.g., European Parliament, 2022; G20, 2019). The value of data is also explicitly and implicitly acknowledged by firms that sell data services, and investors who take dataset assets into account in stock market valuations or mergers and acquisitions.

However, despite the broad recognition of its value, and of the need to develop appropriate policy frameworks to enable data use in societally valuable ways, there is no consensus method for empirically determining the value of data. Without this, the full potential of data will not be realized (Verhulst, 2018). There are not even many examples of markets for data that would indicate a private valuation (although not the wider social value). Yet estimates of the value of data are needed to determine an appropriate level of investment, as well as a better understanding of how data can contribute value to the economy and how to govern the collection and use of different types of data.

This article presents an overview of a range of alternative methods for data valuation, including those proposed in the existing literature. This includes some relatively widely used methods and others that are more specialist or preliminary. However, we do not discuss all the methods discussed in the wider “grey” literature and consultancy reports (see Laney, 2017 for an excellent example). While useful, and placed in the typology for completeness, these are generally either tailored to business strategy, or more qualitative or ad hoc.

Some methods described here (although not all) omit the significant opportunity cost of lack of access to data (European Commission, 2022; Verhulst, 2022). Nor do all of them consider the costs and risks involved in gathering and holding data, such as storage fees, fines or potential reputational damage from data breaches. These certainly need to be accounted for in considering the net benefits of creating and using data. We do not include here some other considerations, particularly estimates of data volume measures or of the stock of data on balance sheets. Finally, we do not consider here the need for complementary investments, or the issue of “missing” data that could improve decision-making. Again, these are an important part of the broader policy evaluation of investment in and use of data but are out of scope here.

2 | PROPOSED DATA VALUATION METHODOLOGIES

The term “data” is broad. We take it to refer to a produced asset consisting of stored digital records structured in such a way that they can be used to inform decisions. Data creates value through its use as information or insight, which will vary in different contexts. The value will also depend on the rights of use associated with the data, which can range from a full transfer of legal ownership of a database, for example, through acquisition of a company, through to a limited license to access (copies of) data or analytics conducted on it. The owners and controllers of databases will generally limit access, both for legal data protection reasons but also because the

TABLE 1 Characteristics affecting the value of data.

Economic lens	Information lens
Non-rival/excludable	Subject
Externalities (positive and negative)	Generality
Increasing/decreasing returns	Temporal coverage
Option value	Quality
High fixed, low marginal costs	Sensitivity
Complementary investments	Interoperability/linkability

Source: Coyle et al. (2020a).

commercial value they gain from it will be enhanced; there is a trade-off between the private and broader social value.

Coyle et al. (2020a) identified key characteristics of data so defined which affect its value-creating potential (see Table 1) and analyzed how the use of data can generate value, including this trade-off. Subsequently, there has been an increasing amount of work on the measurement of data value.

The methods adopted in much of the existing literature can be divided broadly into cost-based, income-based, and market-based methods. These approaches use standard economic concepts and statistics, yet there are drawbacks in each case. In particular, they omit the (potential) wider economic value due to non-rivalry as well as externalities associated with the data (World Economic Forum [WEF], 2021) and the potential negative externalities (especially loss of privacy or data security breaches). Some methods also generate a valuation that does not necessarily account for the value of the content of the data (Ker & Mazzini, 2020). Hence there is also emerging work on less traditional methods which aim to capture the wider economic value of the data including spillovers. The selection of an approach will depend on the purposes for which the data are being valued, as discussed in Section 3 below. For some purposes, the private value to an organization is appropriate, whereas in other cases, including policy development, the broader social welfare value and consideration of trade-offs will be appropriate. Both positive and negative spillovers should be taken into account in this case. In some contexts the cost of lack of access to data may be directly relevant, competition policy being an obvious example, where data holdings are of substantial value to market incumbents but impose a cost in terms of barriers to entry and innovation (Furman et al, 2019). Similarly, concerns about costs in terms of loss of privacy or security breaches may be of direct relevance in considering whether all data gathering and holding is appropriate, even if of private value to the data controller.

Figure 1 below sets out a conceptual typology of approaches, not all of which have generated an empirical literature. It includes the division familiar from economics between revealed and stated preference (what people do in actual economic choice contexts, and what they say they would do in hypothetical choice contexts); and adds two other categories: valuations in terms of impacts affected by data use, and real options modelling. The cost-based, income-based, and market-based approaches form a subset of the possibilities. Below, we discuss the principal methods for which there are some examples of applications in the literature. Others, such as the less formal or quantifiable approaches of stakeholder consultation or customer surveys, are more widely used in consultancy or discussed in the “grey” literature; we do not consider these further here, although they are included in Figure 1 for completeness.

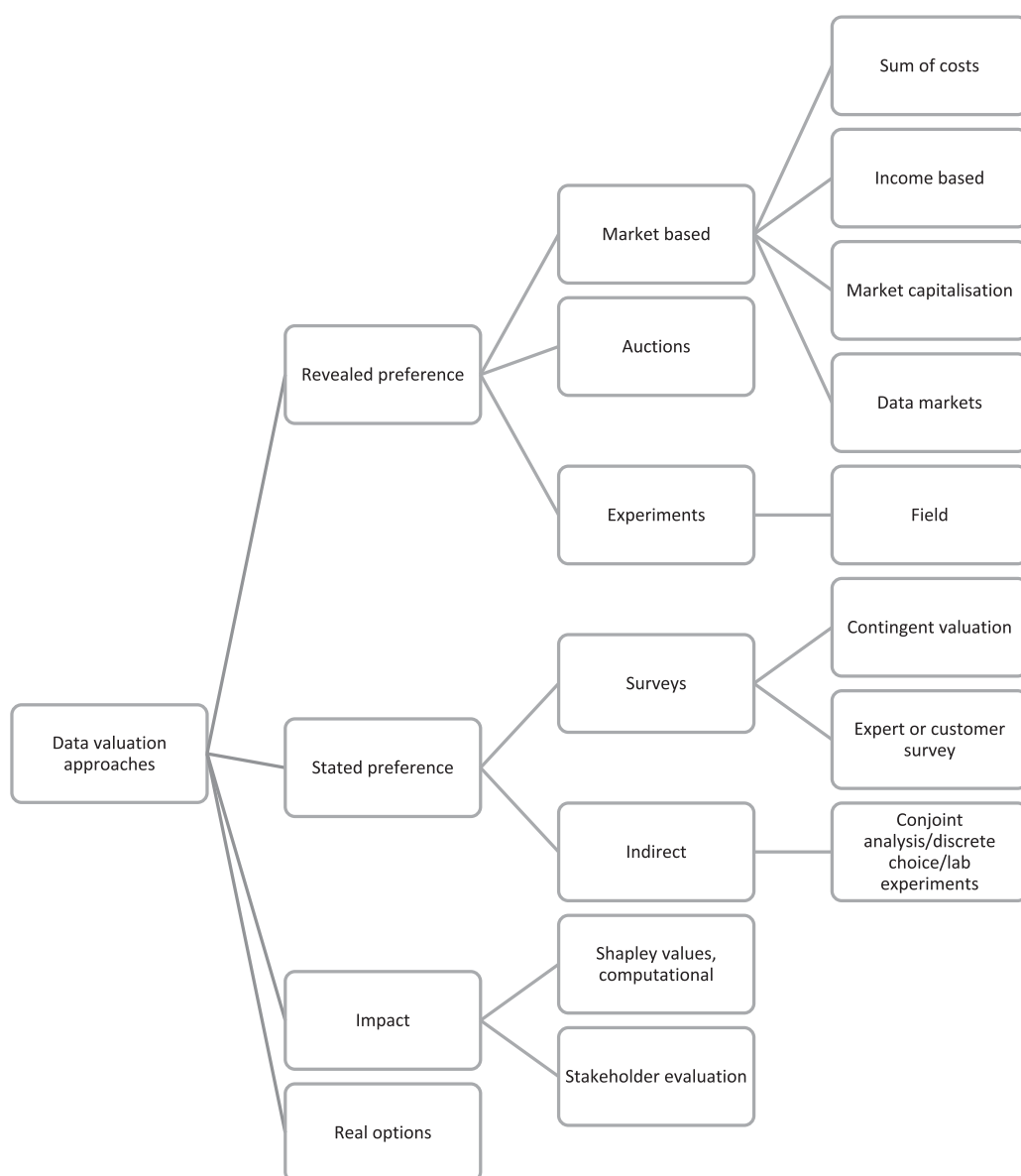


FIGURE 1 Typology of valuation methods.

Most of the existing literature on valuation methods focus primarily on one approach, as summarized in the relevant sections below. However, there also are some comparative pieces of work. These include the Internet of Water taxonomy (2018), Ker and Mazzini's (2020) work for the OECD, and the OECD (2022) supplement to the Going Digital Toolkit.

- Internet of Water is a US-based charity focused on improving water management via improved data infrastructure. It has created a taxonomy of various data valuation methods (2018) they recommend for each of data users, data producers, and data hubs, including applied examples.

- Ker and Mazzini (2020) use four different methods to calculate the value of data using currently available information on data value generation. These are a cost-based approach, an income-based approach, an approach based on market capitalization, and one based on the link between trade flows and data flows.
- The OECD Going Digital Toolkit aims to help countries generate digital development policies. As part of this initiative, a supplementary note (OECD, 2022) stated the importance of estimating the value of data as well as summarizing various case studies of data valuation. It defined content and context as the two crucial components of data value. The note focuses on cost-based approaches including those within the System of National Accounts frameworks, as these are currently the methods being adopted by national statistics offices; but its appendix includes summaries of some other approaches to valuation.

2.1 | Revealed preference: Sum of costs

Cost-based methods are currently one of the most common ways in practice to value data, and likely to remain so. It is the standard practice of the System of National Accounts (SNA) where, if value cannot be directly observed through a market transaction, a “sum-of-costs” approach is recommended as an approximation to exchange value (Ker & Mazzini, 2020). The resulting figure is assumed to give a reasonable lower-bound estimate of the value of the data, under the assumption that an asset will generally have an expected value at least as great as its cost, even taking into account depreciation (an assumption that may not hold in all cases as some data-collection is mandated by law). The approach involves identifying all the costs (including consumption of fixed capital) involved in generating, collecting, storing, and replacing a dataset, as well as the costs to the firm if the data were to be lost or otherwise involve cost (such as compensation for security breaches) (Statistics Canada, 2019; WEF, 2021).

Various extensions of the cost-based approach include the “Modified Historical Cost Method” (MHCM) and the consumption-based method. The MHCM takes the costs and adjusts for data-specific characteristics by giving zero value to duplicated or unused data, then weighting the value by usage rates and accuracy, and setting a depreciation rate based on the purpose of data use (if known). The consumption-based method uses the standard cost-based approach, assuming users value data at least as much as their cost to acquire the data from producers, but adjusts this value to reflect how often the data are downloaded or used by consumers (Internet of Water, 2018).

Because of its relative ease to calculate, and the precedent of using sum-of-costs approaches for other intangible assets, this method has been trialed by a number of national statistical offices (OECD, 2022). These include Statistics Canada (2019), the US Bureau of Economic Analysis (BEA) (BEA, 2019), and the UK Office of National Statistics (ONS) (ONS, 2019), and Statistics Netherlands (2021). These are all similar in their use of labor force surveys and other employment data to determine what proportion of activity in the economy involves data-related tasks, and thus how much is invested in data, with a mark-up for other associated expenditure. Subsequent work by the BEA has used unsupervised machine learning to better categorize occupations and tasks as “data related” to improve this metric (BEA, 2020, cited in OECD, 2022).

Others have also used labor market statistics to calculate the value of data. For example, Goodridge and Haskel (2015) build on earlier work by the ONS to capture a wider range of occupations that work in the data building or knowledge creation sectors. They apply the same approach to these newly identified occupations, estimating that this reveals £1.4 billion of previously unidentified spending on data-related tasks up to a total of £5.7 billion. In contrast, Ker

and Mazzini's (2020) cost-based approach focusses on business statistics databases, considering the amounts spent on data storage products, such as hardware, software, and services by US businesses.

However, national level cost-based approaches rely on having well-classified data at the micro-level. This will be difficult to implement and there are several blurred lines that make classification harder. For example, it is not necessarily obvious how costs for data relative to other costs associated with the related databases might be allocated. It is also difficult to know how accurately occupational surveys can pinpoint the relative investment in different points on the data value chain. Individuals involved in data-related occupations may be responsible for the generation of multiple types of data asset, and may also contribute to software assets (OECD, 2022). Goodridge and Haskel (2015) demonstrate how inadequate precision in occupation selection can omit 20% of annual spending on data, and lead to underestimation of data value in an already conservative approach. On the other hand, some part of the value created by data may be embedded in intellectual property assets created from data, which would be hard to separate and—as IP assets are already included in the SNA framework—would imply double counting.

2.1.1 | Data in the national accounts

The System of National Accounts (SNA) is the United Nations (UN) framework setting “the internationally agreed standard set of recommendations on how to compile measures of economic activity” (United Nations, n.d.). It ensures countries have consistent accounting rules and classifications, making international comparison possible. The classification of data assets within this framework has become increasingly important, as has their valuation. The next update of the SNA is due in 2025, and will include an update on data valuation, among other issues arising from digitization (UN Economic & Social Council, 2022).

The current SNA groups databases along with computer software as a category of intellectual property products within capital formation; data are a produced asset, as opposed to the non-produced underlying observable phenomena. The proposals currently under discussion make a distinction between “long-lived” and “short-lived” data, where “long-lived” data are used in production for over 1 year. This is justified as follows: “When it produces an economic benefit to its owner(s) by using it in production for at least 1 year (2008 SNA, 10.33), termed “long-lived data” in this paper, then data are also to be included in the SNA asset boundary. Short-lived data (i.e., useful for less than 1 year) are instead to be considered as intermediate consumption when it is purchased from third parties, or as the product of an ancillary activity, that is, an integral part of the primary activity, when the production is taking place within the same unit,” (UNSTATS, 2022). These “long-lived” databases would be included in the national accounts, under a newly created fixed asset category of “Computer software, data and databases.” There are currently no proposed changes to the recommended methods of valuation, and several conceptual and practical issues remain unresolved as of writing

In advance of the 2025 update of the SNA, the Digitalization Task Team is still in the process of defining its recommendations for data valuation. The guidance note for the 2025 update that will incorporate data as a produced asset (in addition to the subset of databases already included) recommends using a sum of costs approach for valuing data that is developed on own-account and used internally, while using the market price if the database is developed for sale or license (UN Economic & Social Council, 2022). There are drawbacks to this approach in that there is an inconsistent treatment for data valuations depending on whether the data are developed internally

or for sale in the market (BEA, 2019), and there is also no consideration of the value of the content of the data. Instead, the value of the information contained in the data are considered a non-produced asset, while the market value of the useful information is considered as goodwill in company accounts (Ahmad & van de Ven, 2018). Furthermore, the approach makes no allowances for efficiency of data collection; modern methods of data collection, including official statistics, increasingly involve using online tools that are more cost-efficient than previous survey-based and often even paper-based methods. A measure of real investment in data, incorporating a decline in price, would adjust for this.

2.2 | Revealed preference: Income-based methods

Income-based approaches use the expected revenue streams (such as the sale of marketing analytics or information services) being generated to estimate a value for the underlying data (Coyle et al, 2020a; OECD, 2022; Slotin, 2018). This is obviously most feasible when there are revenue streams directly attributable to use of the data (CGMA, 2012).

One common income-based approach is the “relief from royalty” method, which is often used for other intangible assets such as trademarks and patents. This is a direct (revealed rather than surveyed) measure, as it indicates how much money a company can save by creating or owning the data rather than paying royalty payments to license it (Ibid).

Ker and Mazzini (2020) also use an income-based method by calculating the reported revenues related to compiling and selling data, based on business survey data. However, current industry data classifications do not well identify these cases, so there are limited opportunities for applying this method. In addition, by identifying only the cases where data are the “product,” this method excludes the value created by data in “data-enhanced” firms that do not sell data directly but use it to improve their products or processes or generate revenues from data analytics (e.g., Apple and Google, or manufacturing firms using data to better manage their supply chains). Similarly, monetization via data sharing arrangements among firms—as previously practiced by Facebook making data available to other technology companies via APIs (New York Times, 2018)—would be omitted.

Income-based approaches involve more judgment than cost-based approaches. This is because data costs are relatively well defined and are not affected by the insights generated from the data. Revenue streams can include revenue streams enhanced by data insights but also other inputs. Therefore, income-based valuations may have to estimate a counterfactual income stream. In addition, it is often very hard to predict the value of a revenue stream before those gains are realized (OECD, 2022), making it a retrospective method, and difficulties arise when attempting to identify whether the revenue has come from the data or another contributing factor (CGMA, 2012).

2.3 | Revealed preference: Market-based methods

Market-based methods use observable prices. When they are available, market prices are always a preferred method for valuation—although as noted earlier, market prices will be only a partial estimate of total social value (Coyle & Diepeveen, 2021). However, there are relatively few observable market prices for data. Most data are currently created and used internally by firms or other organizations rather than being sold or traded. Even well-established data markets

such as credit scoring generally do not post their prices—although revenues from sales of credit scores are reported in company accounts of credit bureaux, enabling some inferences to be made about average prices (Soloveichik, 2022). It is worth noting that lenders pool their own data by providing it to the credit scoring bureaux, which then sell services back to the lenders, hence demonstrating that the value of the aggregated data is greater than the value of the sum of its parts.

The academic literature looking at market-based approaches to data valuation can be split into three categories: those advocating the use of data marketplaces and looking at desirable characteristics of data pricing; those which use the market capitalization of firms to calculate value; and those using global data flows to elicit a valuation.

2.3.1 | Data marketplaces

There is a growing literature on pricing in data markets and exchanges, and the potential for well-functioning data marketplaces. Data marketplaces are proposed as a mechanism to increase the amount of value generated from each dataset by minimizing transaction costs, generating transparency around pricing, and exploiting data's ability to create value for multiple agents simultaneously (Heckman et al., 2015a). There are a growing number of attempts to create data marketplaces, with varying degrees of success (Da Silva & Nunez, 2022).

Pei (2020) offers a survey of the literature on data pricing, with insights into how data are typically sold in a competitive market. Pei notes that data suppliers often bundle data together, price according to what consumers will pay (as opposed to setting a mark-up over cost), and also commonly engage in price discrimination. They attribute this to data being non-rival, with marginal costs close to zero, and price discrimination being relatively easy to implement. The paper mentions some potential methods for setting data prices, but in general reflects the primarily theoretical nature of the existing literature with a focus on desirable characteristics rather than practical valuation methods.

In contrast, Da Silva and Nunez (2022) discuss data valuation primarily in relation to the creation of data marketplaces. They discuss case studies in China, New Zealand, the European Union, and Colombia, where there have been efforts to create data marketplaces or otherwise encourage data exchange. These endeavors reflect a policy interest in pricing data, with the report going as far to say: "The success of initiatives to create data marketplaces depends largely on pricing data" (Da Silva & Nunez, 2022, p.35). They also mention a number of barriers to marketplace development, including heterogeneous and restrictive regulations on data, and a lack in trust, leading to low participation.

Koutroumpis et al. (2020) review the literature on markets for data, including its institutional history, the wider literature on markets for ideas, and the main data market matching mechanisms, with examples. They note that data are typically an intermediate good and an experience good, meaning its value is not observable before consumption. They summarize and provide examples of several types of data marketplace categorized by the matching mechanism, described in Table 2 below. Although not discussed in the paper, the growing number of specialist data companies could also have been included here, including long-established producers such as Nielsen (audience data) and a burgeoning number of new "big data" providers selling satellite or tracking data for example. Koutroumpis et al. conclude that it is only possible to achieve either large markets with little control or small markets with more, albeit not full, control.

They argue this is because both trust in the data and in contract enforcement preventing data use beyond the contract terms are key to achieving market "thickness" and a self-sustaining and

TABLE 2 Types of data marketplaces and their characteristics.

Matching	Marketplace design	Terms of exchange	Liquidity	Transaction costs	Safety	Examples
One-to-one	Bilateral	Negotiated	Low	High	High	Personal data brokers and creators, Acxiom
One-to-many	Dispersal	Standardized	High	Low	Low	Twitter API, Facebook API
Many-to-one	Harvest	Implicit barter	High	Low	Variable	Google Waze, Google Search
Many-to-many	Multilateral	Standardized or negotiated	High	Low	Low	None

Source: Koutroumpis et al. (2020).

efficient market. Moreover, there are inherent challenges in establishing markets for data. Markets depend on some reasonable standardization of the goods being sold and sufficient information about their quality, whereas there are no standard units for data in general, and there is an inherent asymmetry of information: it is an “experience good” whose quality is unknown until it has been acquired. It is well known that markets are prone to collapse when such asymmetries of information exist.

Nevertheless, a number of authors have proposed theoretical approaches to pricing in data marketplaces. Deep and Koutris (2016) present a general information-theoretic framework for price functions applying to database queries that prevent arbitrage in data markets. Agarwal et al. (2019) propose a design for a two-sided data marketplace for training data, along with an algorithm that would enable prices to be set in this marketplace.

The record of actual many-to-many data markets is slender. The Microsoft Azure DataMarket was set up in 2010 with ambitions to be a “one-stop shop for premium data and applications”, which would allow companies to easily monetize their pre-existing datasets (Staten, 2010). However, a lack of customer interest (market thickness), led to its shutdown in 2017 (Ibid.). Other examples of data markets currently being trialed include the Shanghai Data Exchange and the Ocean Market. The Shanghai Data Exchange launched in November 2021, with 20 data products sold by Shanghai companies—including three major telecom operators (Protocol, 2021). They are heavily regulated, and the market is limited to one-to-one transactions only (Financial Times December, 2021). The Ocean Market, founded in 2017, more closely resembles a competitive market, and uses an Automated Market Maker to generate data prices—a protocol typically used to automate price calculations in decentralized finance, such as the initial price offering on a new cryptocurrency. Datasets sold here range from 6000 Indian food recipes, to Reddit threads annotated ready to be analyzed via natural language processing (Ocean Protocol, 2022).

2.3.2 | Market capitalization-based methods

One way to value data which is primarily used by a firm is to look at the impact of the data on the firm’s market capitalization. The approach is imperfect because market capitalization depends on multiple assets and capabilities but is appealing in being grounded in observed market prices. Ker and Mazzini (2020) use the entire market value of the firm for a limited set of firms, whilst Li and Chi (2021) and Coyle and Li (2021) look at organizational capital as a separate contributor to market capitalization.

Ker and Mazzini's approach is based on the market capitalizations of "data-driven" firms. These are firms with business models such that their value is "primarily" derived from their data and analytics. In Ker and Mazzini's estimate, these are identified as those featuring in lists such as "The Cloud 100 list" or "The Top 100 Big Data companies" and the method is then applied to those based in the United States (US) with listings on the NASDAQ or the NYSE. This includes smaller firms involved in data management and analytics alongside more high-profile examples such as Amazon, Google, or Facebook. Ker and Mazzini estimate that "data-driven firms" in the US are worth over 5 trillion US dollars.

Coyle and Li (2021) exploit empirical relationships found in Li and Chi (2021) on the impact of data-driven disruption on the organizational capital of incumbents to estimate the size of markets for data by sector, as well as their annual growth. To demonstrate the method, they measure the impact of a new data-driven entrant, AirBnB, on the non-digital incumbents in that market, such as Marriott, and find that the depreciation rate of the organizational capital of non-digital incumbents increased (organizational capital is estimated from company accounts using a standard method). This is attributed to the incumbent's failure to use data and thus a relative decline (compared to data-driven newcomers) in the value of the firm-specific knowledge. The value of lost organizational capital is then interpreted as the amount the company should be willing to pay to acquire and use data. The method can be aggregated across companies to give sector-wide estimates.

2.3.3 | Data flows

Data flows have a significant advantage in sometimes being observable in a market where there is little information. There is a strong correlation between the volume of data flows and the value of data in dominant online platforms (Li & Chi, 2021). Therefore, there have been some attempts to value global data flows. These form a subset of market-based measures but are limited in only covering data that crosses international borders. However, both Ker and Mazzini (2020) and Coyle and Li (2021) consider but dismiss the link between volume of data flows and the emergence of data value *in a given location*. This is due to the structure of the international technology industry, with large data hubs in certain locations that service a large area, and the frequent need for the value generation from data to be done with local knowledge that is difficult to relocate. The geographical distribution of data value therefore stays the same even if the data storage is moved across international borders.

Ker and Mazzini look at data flows in terms of the volume traded, and the value of the products which are delivered via data flow. They criticize the former approach, as the content of data flows is a far larger determinant of its economic value than its volume. For example, video streaming will typically create a larger amount of internet traffic than e-commerce transactions relative to the economic activity it underpins. Ker and Mazzini instead mainly focus on the value of products delivered through data flows. Here they note the sensitivity of their estimates to the definition of "digitally deliverable products", observing that it will likely be several years before enough countries will record digitally-delivered trade flows for this approach to be practical. They therefore generate estimates using trade statistics in products which are *likely* to be delivered through data flows, identifying the United States to be the largest global hub for data flow trade, with several European nations also showing large amounts of trade underpinned by data.

2.4 | Stated preference approaches

There is a large literature on using stated preference approaches as a method of estimating values in contexts where there are no market prices available. The method involves surveying individuals to elicit their willingness to pay (WTP) for a product or service or willingness to accept (WTA) its loss. A key aspect of these methods is that they can be used to calculate the consumer surplus—that is, the whole of the value consumers attribute to a product given their preferences. This means that the value generated from externalities or the public good aspects of data can potentially be taken into account. (Coyle & Diepeveen, 2021).

These methods are widely used when market prices do not exist, or when market prices do not reflect the full economic welfare of an asset due to externalities, in areas such as environmental and cultural economics. Contingent valuation methods, which ask survey participants to state what they would be willing to pay or accept, have been used previously in some instances of data valuation (BEIS, 2019; Deloitte, 2017; Miller et al., 2013. See also examples in Bennett Institute for Public Policy, 2019b). There has been a vigorous debate in economics about the merits of stated preference surveys (e.g., Hausman, 2012) but there are no obvious alternatives, and much effort has been put into the technicalities of survey design and incentive compatibility (e.g., Haab et al., 2020).

Pilots involving indirect survey methods, such as conjoint analysis or discrete choice analysis, have been trialed for public sector data (Coyle & Manley, 2021; ONS, 2021, 2022). Widely used in marketing, these involve varying the attributes of the product including price to determine which combination respondents would most prefer. These approaches involve the comparison of different attributes of data sets, with the inclusion of price providing a monetary reference point.

The Office for National Statistics (ONS) pilot (2021) used a discrete choice analysis, also known in marketing as choice-based conjoint analysis, in a small survey of ONS employees, to evaluate the additional value of having official data over non-official data. The ONS study combined a survey using pre-existing marketing software with qualitative interviews with the participants as they complete it. Coyle and Manley (2021) applied the same method, without the follow-up qualitative surveys, to explore the value of public sector data being open access. Specifically, they looked at the willingness to pay (among a survey of about 400 economists) for the World Bank development indicators dataset. Both papers discovered some difficulties in using discrete choice analysis, specifically in cognitive overload of participants who are unused to having to quantify the value they place on data and its different attributes. There is a trade-off between surveying participants who commonly use data and therefore understand how data can be valuable for them, and surveying the wider population, which would be more representative but with a higher chance of poor quality or inconsistent responses.

Where data are unlikely to ever be traded, such as in the case of official public statistics, stated preference methods will continue to be useful measures of economic value in debating how much to invest in collecting and maintaining data, what aspects are most useful to users, and potentially how to price freemium access models if necessary or required. As a public good, stated preference methods are suitable for estimating the value of data beyond the private value to data controllers or organizations.

2.5 | Impact-based methods

Impact-based methods identify the value of data as the causal effect it has on outcomes; after all, the purpose of collecting and using data are to improve insight and make better choices (WEF, 2021). Impact-based measures are described by Slotin (2018) who deems them promising for policymakers. This is because of their ability to promote storytelling and directly link data to outcomes, which Slotin argues is likely to be a more persuasive demonstration of value than previous quantitative methods.

Impact-based methods typically exploit natural experiments, such as staggered timings in data availability (Slotin, 2018), or simulating counterfactual outcomes (Arrieta-Ibarra et al., 2020). Slotin (2018) reviews five types of data valuation methodologies, concluding that impact-based approaches are preferable as they are generally easier to understand and to communicate.

Arrieta-Ibarra et al. (2020) provide an empirical study valuing the outcomes of using data in machine learning algorithms. They use counterfactual simulations which vary the amount of data available to machine learning algorithms to compare the effect on company profits. They apply their method to Uber, calculating an upper bound estimate of the value of Uber's data, estimating that data use generates up to 47% of Uber's revenue. If the drivers were fully compensated for the data they created in this upper bound scenario, it would translate to payments of 30 US dollars per driver per day just for their data generation services.

The Internet of Water (2018) framework also recommends impact-based methods for data users, and suggests a "Business Model Maturity Index" as a way to calculate it. This method weights outcomes that have been influenced by data by the relative contribution of the data to the desired outcome. Consider a firm trying to lower costs, with access to energy price data—its choice of energy supplier may be almost fully informed by this data, and therefore the contribution of the data to cost savings via changing energy supplier is high. However, the firm may also use energy price data to inform their decision on which new servers to buy, and whether they should be energy efficient. In this case, the energy price data are one of several considerations, and the weight of the contribution of the data to subsequent cost savings is lower. An extension of this method is the decision-based valuation method which adjusts the value of the data for frequency, accuracy, and quality before weighting the outcomes by the relative contribution of data to the decision. The focus on relative contributions specifies that value comes from improvements to the decision-making process, and therefore should take possible substitutes for using the data into account. However, this does mean the process involves judgment.

Impact-based methods can be highly effective at describing the economic value of data in use, but require a reliable counterfactual. Natural experiments should be exploited where they arise, but these will be limited cases, and context specific. Simulated or estimated counterfactuals have the potential to be highly subjective and can both over- and under-estimate the impact of data on decisions and outcomes.

2.5.1 | Data Shapley values and other computational approaches

Some impact-based methods depart from the approach of valuing data based on the societal value of the application of data-driven insights to focus on private, organizational benefits, and a growing number of computational methods have been proposed in the machine learning literature.

In computer science, Shapley values are used to observe the expected impact of a single observation or unit of information on the quality of a model, relative to a predefined performance metric.

Shapley values are derived from game theory as the unique payoff solution within a public good game that satisfies requirements of group rationality, fairness, and additivity (Jia et al., 2019). In a data context, this is typically used to judge data quality, and to pick out mislabeled data and bad data sources that negatively affect model performance (Ghorbani & Zou, 2019), or for optimal feature selection (Fryer et al., 2021). However, Shapley values could be used to calculate compensation for data provision at a data-point level; in other words, in calculating how much providers of the data—such as users of a digital platform—should be compensated for this provision. The idea is related to interest in the idea of “data as labor”, rewarding individuals financially for the provision of their data to digital platforms (Arrieta Ibarra et al., 2018). The value of individual data points can also help identify the most valuable types of data for further collection (Ghorbani & Zou, 2019).

However, there are some drawbacks, particularly the computational expense of calculation for all data points for all datasets, and the translation of data values into monetary compensation. The computational expense is a key limitation of this method, though the widespread use of Shapley values for other purposes has meant there is already substantial research into more efficient calculation algorithms (e.g., Ghorbani & Zou, 2019; Jia et al., 2019). As a valuation method, it is also important to note that the Shapley value is linked to a payoff function that is the performance metric of the model, rather than a monetary value (Fryer et al., 2021). Therefore, Shapley values are likely to provide a weighting system for compensation, but it will not resolve controversies around the level of compensation needed for data supply.

Finally, Fryer et al. (2021) note that Shapley values are not the only solution to a public good game, but merely the unique solution that satisfies group rationality, fairness, and additivity. These are by no means the only desirable axioms to satisfy, and there are a variety of other solutions which may have more desirable properties.

An alternative approach based on reinforcement learning is demonstrated by Yoon et al. (2020). They propose a data value estimator that is trained using a reinforcement signal from a training data set based on how likely each data point was used in creating the predictive machine learning model. Their aim is to find a value for individual data points; they argue that their approach is less computationally complex and costly than the use of Data Shapleys.

2.5.2 | Stakeholder-based methods

The stakeholder approach is designed to take the value of data to a variety of stakeholders into account. This means that, rather than focusing on only the value of the data to the firm and its profits, employees, customers, supply chain providers, and the wider public should be taken into account. This is a wider definition of value and may include value upstream or downstream omitted from other methods; it can encompass the non-rival aspect of data. For example, the data consultancy Anmut is one that has developed this method and provides a case study of their valuation of Highways England data (Anmut, n.d.).

The WEF white paper (2021) highlights that different organizations are defining value differently, ranging from full social value, to “stakeholder” approaches such as that used by Anmut, and “impact-based” approaches, often favored in development sector organizations. They conclude that the stakeholder approach is the preferred method for evaluating data ecosystems. They specify a broad range of stakeholders, “including businesses that collect and process it, partners and third parties that they share it with, governments, customers, employees and especially gig work-

ers, and individuals and communities who provide and are impacted by the data,” and encourage the consideration of environmental, societal and governance (ESG) factors as sources of value. The inclusion of multiple stakeholders explicitly within the valuation process addresses the aim of evaluating the broader social value of data. However, this method also involves matters of judgment.

2.5.3 | Impact on growth and jobs

It is worth noting that in the broader grey literature, impact-based approaches may also be used to capture some of the wider economic value generated by data, for example, via higher growth and more jobs in a sector that has embraced data.

For example, the Progressive Policy Institute (Mandel, 2017) looks at the “app” economy in the US. They note that, whilst global app revenues reached 45 billion US dollars in 2016, this is unlikely to reflect the true value of the app economy, as many free apps, such as Instagram or Google search, generate no incomes via an app store. Their preferred method is to use employment metrics, specifically job postings requiring “App Economy-related skills, such as knowledge of iOS or Android”. They find that there were 1.7 million jobs in this sector in the US in 2016, with an annual growth rate of 30% between 2011 and 2016. However, the paper does not directly measure the wages of those working in this sector or discuss the value of the data asset they create.

2.6 | Real options analysis

Several of the above approaches, notably the impact-based and stakeholder-based methods, are best suited to calculate value *ex post* the use of a dataset. However, an estimate of data value would often be useful in advance. Because it is non-rival, the full extent of the data’s use cases does not need to be defined at the point of data collection. Instead, data can be worth collecting due to its *potential* use cases. This can be thought of as the option value, or the “right but not the obligation” to generate insights from data in the future. The option value of data has been implicitly acknowledged by firms, who often collect data for unknown future purposes (Coyle & Diepeveen, 2021). In these cases, firms can wait for more information (e.g., information about policy changes, technological advancement, changes in tastes, etc.) to see if data are worth processing. Firms will only process and analyze the data if the expected value of the insights generated would be material but the act of collecting the data does not necessarily oblige them to do so. This suggests a real options approach could be useful.

Real options were first mentioned by Myers (1977) as a parallel to financial options. They represent the value of flexibility within a project that allows firms to delay committing to certain decisions until more information is available or uncertainties have been resolved. This means they are able to capitalize on upside opportunities without being exposed to downside risk (de Neufville et al., 2006). Real options value is generally calculated using simulations of future scenarios whereby the firm has taken the optimal action in each scenario (in this case, acquiring, processing, and analyzing the data, or not). The value of the data in each scenario is necessarily non-negative, since the firm can always choose not to use it. The expected value of the (data) asset over all scenarios is then taken, generating the real options value of the data.

One common method of calculating real option value is the Black-Scholes formula from finance. This is the approach proposed by Coyle and Diepeveen (2021), and used by Coyle, Gamberi and Manley (in progress). The main benefit of this method is its simplicity, with only five



FIGURE 2 Data value chain.

parameters needed to calculate data value, though there remain difficulties in practice. Cheaper and more accessible computing power has made it easier to calculate real options value using simulation methods, including Monte Carlo simulation, with the accuracy of the simulations only due to improve as more datasets are systematically valued and the dynamics of data value are better understood.

3 | DISCUSSION

One suggestion in response to the absence of a single consensus or best method for valuing data are to propose a systematic way of linking methods selected to different data types and purposes. Heckman et al. (2015b) for example propose that data should be algorithmically classified to determine which method should be applied, and valuations should be externally validated by surveys.

Some key considerations for determining the most appropriate method which have been mentioned in this part of the literature are:

- *what is being valued?*
- *who is valuing the data?*
- *when is the valuation taking place?*
- *what is the purpose of the valuation?*

For each problem, there may also be a valuation methodology which would be wrong (CGMA, 2012). It is interesting to apply these considerations to the valuation methods described earlier.

3.1 | What is being valued?

There are a variety of things that could be referred to as “data.” The possible distinctions are illustrated in the “data value chain,” shown in Figure 2, which sets out different stages from the generation of raw data up to the decisions made using data insights generating the potential end-user value.

In general, the raw data are of least interest, and some of the literature goes as far as to state that raw data does not hold any value on its own (e.g., ODI, 2018; WEF, 2021). Observables, as noted above, are not considered as produced assets. Even with cost-based methods, in many ways the most straightforward approach, it is often almost impossible to distinguish between costs associated with raw data generation and database formation (OECD, 2022).

Many of the methods discussed above, including income-based, impact-based, and stakeholder-based methods, explicitly refer to the value of insights, decisions, and actions enabled by the data, or the final stages in the data value chain. They implicitly bundle the value of data with the value of the analytics applied to it and the translation of insights into decisions. Data in the hands of

a dynamic firm (or accessed by others) is therefore more valuable than the same data in a more static firm. For example, Brynjolfsson et al (2021) find that productivity is significantly higher among manufacturing plants that use predictive analytics in general, but some firms can gain no benefits at all. They find that productivity gains seen are almost exclusively within firms that have high levels of IT (informational technology) capital, educated employees, or production process design which favors high flow efficiency.

Some valuation methodologies, notably cost based approaches, result in a value for a dataset that can be entirely independent of the content of the data, the context of its collection, and the use to which it is put. Given the importance of the content of the data, and the importance of the context for trust in the data (WEF, 2021), they are likely to form incomplete estimates.

Finally, the insights generated from data can potentially be generated from samples rather than the whole of a dataset. When costs are also taken into account, the question for generating value becomes “what is the minimum data needed for good decision-making?”

3.2 | Who is valuing the data?

The perspective of the person or entity valuing the data will affect the appropriate method. The Internet of Water (2018) classification goes so far as to break down methods along these dimensions, looking at the methods most appropriate for data producers, data users, and data hubs. The public sector is also likely to consider data value differently to private agents.

Data producers in the public sector are likely to focus more on the costs of data generation, given their budget and statutory constraints, and may be more focused on the value of the raw data and the managed database than on the insights from the data. They may therefore opt for cost-based approaches to valuation. The difficulty is finding a way of aggregating the vast array of data being produced in the economy when the use values are even more heterogeneous than is the case with other economic assets. Increasingly, public sector bodies are being asked to monetize data they hold, but the monetary value in terms of potential revenues will differ from the potential public value of open access.

Private sector data producers are likely to focus on the impact of using the data as well as costs. As noted, a single data set can generate many different decisions and outcomes depending on context.

Data users are likely to focus more on the insights gained from analyzing the data than the specifics of the underlying data. Methods that value the insights and impact of the datasets may therefore be more appropriate. What is being valued here may also depend on the state in which the users receive the data (whether bundled with analytics or in rawer form). Users are also most likely to consider the option value of data as they consider future applications.

Data hubs, intermediate agents between producers and users, are likely to focus more on the market price for their data. Established data hubs are likely to have historical data on realized sale prices, although as noted earlier these are likely to be differentiated among users, not being posted in open markets or exchanges.

The public sector is likely to focus on data value from a wider societal view. Further research is needed for this perspective, as current empirical work has tended to focus on private valuations, which exclude the value of the externalities generated by data (Coyle & Diepeveen, 2021) and may ignore some of the more indirect or unanticipated revenue streams (Slotin, 2018). This tendency for undervaluation is even greater when in the context of open data (Deloitte, 2017; WEF, 2021).

Slotin (Ibid.) argues that top-down macroeconomic estimates of value, which would be more likely used by the public sector, typically overestimate the value of data by ignoring the potential substitutes a firm could use. It is important to note that the value of data to an organization is affected by regulatory changes, by legal considerations, and also by the emergence of alternatives that make the data held non-exclusive. The aggregate methods assume the economy is competitive (and hence market prices are reflective of allocative value) but non-competitive prices include elements of monopoly rent. Such rents can be eroded by new entrants or alternative data sources due, for example, to technological advances (such as the proliferation of new sources of satellite imagery).

However, the growing literature about the impact on productivity of data use by firms (e.g., Brynjolfsson et al., 2021; Bajgar et al., 2021; Coyle et al., 2022) suggests there are not good substitutes for data as an input into production decisions. Data is an intangible asset that enables firms to gain a productivity advantage and to grow (in increasingly concentrated markets) because its use requires complementary skills and capabilities in the firm. Firms able to capture monopoly rents from data will attribute higher value to it than potential alternative users, but their private valuation (or their resulting market valuation) will diverge from the social value.

3.3 | When is the valuation taking place?

There is an important distinction to be made between methods which are appropriate for *ex ante* data valuation, and those which are for *ex post* valuation.

Ex ante valuations are limited because the insights, actions and impacts made possible by data are highly uncertain in prospect. The option value of data is significant. Few methods discussed here incorporate forward-looking values, in contrast to the valuation of most economic assets.

Ex post valuations provide more information with more certainty. This allows for more methods to be used but raises the question of the extent to which different types of data depreciate, and how quickly. In some contexts (such as data for autonomous vehicles or financial markets) real time data are valuable and depreciation is extremely rapid, whereas other types of data (such as map or other reference data) may depreciate slowly if at all.

3.4 | What is the purpose of the valuation?

The reason for wanting to value data may help determine which methodologies are appropriate.

For example, cost-based methodologies are preferred for generating a cautious but reasonably certain estimate, while market-based approaches are able to generate a figure justified by the choices people actually make in the course of transactions. Nevertheless, these methods omit some of the known aspects of what makes data valuable when it is used—in different decisions or by different people. A key justification Slotin (2018) gives for impact-based measures is their potential for storytelling and thus linking data to tangible outcomes. Stakeholder approaches also enable holders and users of data to appreciate the value of the data assets, which might implicitly have been undervalued previously (Anmut, n.d.).

It is worth underlining again the gap between private and public purposes, with implications for approaches to valuation. Private sector data producers and users will have an outcome

TABLE 3 Criteria for selection of valuation method.

What is being valued	Who is valuing	When is it done	What is the purpose
Data:	Private entity:	Before decisions:	Generating income:
<i>Sum of costs</i>	<i>Sum of costs</i>	<i>Option value</i>	<i>Income based</i>
<i>Market based</i>	<i>Income based</i>	<i>Stated preference eg discrete choice</i>	<i>Impact eg Shapley, narrative</i>
	<i>Market based</i>		
	<i>Impact</i>		
Analytics:	Public entity:	Evaluation/accounting:	Policy development:
<i>Sum of costs</i>	<i>Stated preference</i>	<i>Cost based</i>	<i>Stated preference eg contingent value</i>
<i>Income based</i>	<i>Impact eg stakeholder survey</i>	<i>Income based</i>	
<i>Market cap</i>		<i>Impact</i>	<i>Impact eg stakeholder survey</i>
<i>Impact eg stakeholder evaluation</i>			

or impact they hope data use will improve, such as productivity or profitability. Data-holders with market power will place a higher value on their data than if it were valued in a competitive environment and available to other companies. Other wedges between private and social value are due to the non-rival character of data. Public bodies absorb the costs of data investment and stewardship of official data holdings, as do other open data providers. There is growing debate about the role of data intermediaries, and about the allocation of value to the individuals who ultimately provide their data and thus create value. The data-as-labor proposal (Arieta Ibarra et al., 2020) mentioned earlier is one individually-oriented approach; another example is the creation and allocation of value through data co-operatives (Dawande et al., 2022).

Examples of applying these considerations are summarized in Table 3.

4 | CONCLUSIONS

The above four key considerations go beyond the scope of this survey and assessment of practical applications of data valuation methods. There is a growing need to be able to value data, to both encourage data investment and use, and to understand the impact of data on the economy, enabling productive and socially valuable uses. However, there remain relatively few robust empirical approaches, and certainly no consensus about which to apply for different purposes. Existing valuation methodologies for data all have drawbacks. There is both a need and an opportunity to further develop methodologies able to empirically capture the economically valuable characteristics of data, both the private value and—for policy purposes—the wider social value.

ACKNOWLEDGMENTS

We are very grateful to Sylvie Delacroix, Sumedha Deshmukh, Stephanie Diepeveen, Vilas Dhar, Luca Gamberi, Owen Garling, Sam Gilbert, Verity Harding, Julian Huppert, Jared Keller, Neil Lawrence, Derek McAuley, Claire Melamed, Maria Savona, Jeni Tennison, Olivier Theroux, and Stefaan Verhulst for their helpful comments and to many of these colleagues for their contribu-

tions to a Bennett Institute workshop on data valuation. Anonymous reviewers provided helpful comments. We also thank the Omidyar Network for funding this work and particularly Sushant Kumar for his comments and intellectual support.

DATA AVAILABILITY STATEMENT

N/A.

ORCID

Diane Coyle  <https://orcid.org/0000-0001-7243-1641>

REFERENCES

- Agarwal, A., Dahleh, M., & Sarkar, T. (2019). A marketplace for data: An algorithmic solution. In *ACM EC '19: ACM Conference on Economics and Computation (EC '19), June 24–28, 2019*. ACM, New York, NY, USA. <https://doi.org/10.1145/3328526.3329589>
- Ahmad, N., & van de Ven, P. (2018). *Recording and measuring data in the System of National Accounts*. Paper prepared for a meeting of the Informal Advisory Group on measuring GDP in a digitalised economy. 9 November. Paris, France. OECD. https://unstats.un.org/unsd/nationalaccount/aeg/2018/M12_3c1_Data_SNA_asset_boundary.pdf
- Anmut. (n.d.). Data valuation | Why it matters & how it's done. <https://www.anmut.co.uk/an-introduction-to-data-valuation/>
- Arrieta-Ibarra, I., Caruso, T., Hernández, D. J., & Weyl, E. G. (2020). *An empirical study of the value of data*. Mimeo.
- Arrieta Ibarra, I., Goff, L., Jimenez-Hernandez, D., Lanier, J., & Weyl, G. (2018). Should we treat data as labor? Moving beyond “free”. *AEA Papers and Proceedings*, 108, 38–42.
- Bajgar, M., Criscuolo, C., & Timmis, J. (2021). Intangibles and industry concentration: Supersize me. [OECD Science, Technology and Industry Working Papers No. 2021/12]. <https://doi.org/10.1787/ce813aa5-en>
- BEA. (2019). *Treatment of data in National Accounts*. Paper prepared for the United States Bureau of Economic Analysis. <https://www.bea.gov/system/files/2019-05/Paper-on-Treatment-of-Data-BEA-ACM.pdf>
- BEA. (2020). *A labour costs approach using unsupervised machine learning*. OECD Publishing.
- BEIS. (2019). Companies house data: Valuing the user benefits. <https://www.gov.uk/government/publications/companies-house-data-valuing-the-userbenefits> on 21 August 2020
- Brynjolfsson, E., Jin, W., & McElheran, K. (2021). The Power of prediction: Predictive analytics, workplace complements, and business performance [Author's Final Manuscript] <https://dspace.mit.edu/handle/1721.1/137842.2>
- Coyle, D., Gamberi, L., & Manley, A. (2022). Real options valuation for datasets (in progress).
- Coyle, D., & Diepeveen, S. (2021). Creating and governing social value from data [Conference Presentation]. IARIW-ESCoE Conference ‘Measuring Intangible Capitals and their Contribution to Growth’, RSA House. https://iariw.org/wp-content/uploads/2021/10/diane_paper.pdf
- Coyle, D., & Li, W. (2021). The data economy: Market size and global trade [ESCoE Discussion Paper No. 2021-09]. <https://www.escoe.ac.uk/publications/the-data-economy-market-size-and-global-trade/>
- Coyle, D., & Manley, A. (2021). Potential social value from data: An application of discrete choice analysis [ESCoE Discussion Paper 2021-17]. <https://doi.org/10.2139/ssrn.3973036>
- Coyle, D., Diepeveen, S., & Wdowin, J. (2020a). *The value of data summary report*. Cambridge: The Bennett Institute for Public Policy. <https://www.bennettinstitute.cam.ac.uk/publications/value-data-summary-report/>
- Coyle, D., Lind, K., Nguyen, D., & Tong, M. (2022). Are digital-using UK firms more productive? [Discussion Papers ESCoE DP 2022-06, Economic Statistics Centre of Excellence (ESCoE)]. <https://www.escoe.ac.uk/publications/are-digital-using-uk-firms-more-productive/>
- Da Silva, F., & Núñez, G. (2022). The era of platforms and the development of data marketplaces in a free competition environment [Project Documents (LC/TS.2021/173)]. Economic Commission for Latin America and the Caribbean (ECLAC). https://repositorio.cepal.org/bitstream/handle/11362/47773/S2100763_en.pdf?sequence=1&isAllowed=y

- Dawande, M., Sameer, M., & Liying, M. (2022). Robin hood to the rescue: Sustainable revenue-allocation schemes for data cooperatives [Author's final manuscript]. <https://doi.org/10.2139/ssrn.4020556>
- Deep, S., & Koutris, P. (2016). "The design of arbitrage-free data pricing schemes," CoRR, vol. abs/1606.09376. <http://arxiv.org/abs/1606.09376>
- De Neufville, R., Scholtes, S., & Wang, T. (2006). Real options by spreadsheet: parking garage case example. *Journal of Infrastructure Systems*, 12(2), 107–111.
- Deloitte. (2017). Assessing the value of TfL's open data and digital partnerships. <http://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf>
- European Commission. (2022). Impact assessment report and support studies accompanying the proposal for a Data Act. <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-report-and-support-studies-accompanying-proposal-data-act>
- European Parliament. (2022). Governing data and artificial intelligence for all: Models for sustainable and just data governance. [https://www.europarl.europa.eu/stoa/en/document/EPRS_STU\(2022\)729533](https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2022)729533)
- Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. <https://arxiv.org/pdf/2102.10936.pdf>
- Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019) Unlocking Digital Competition. Report of the Digital Competition Expert Panel. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf
- Ghorbani, A., & Zou, J. (2019). Data Shapley: Equitable valuation of data for machine learning. *Proceedings of the International Conference on Machine Learning*, PMLR, pp. 2242–2251.
- Goodridge, P., & Haskel, J. (2015). How much Is UK business investing in big data? [Discussion Paper 2015/05]. Imperial College London. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.926.243&rep=rep1&type=pdf>
- G20. (2019). Osaka Leaders' Declaration. https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/en/documents/final_g20_osaka_leaders_declaration.html
- Haab, T., Lewis, L., & Whitehead, J. (2020). State of the art of contingent valuation. In *Oxford Research Encyclopedia of Environmental Science*. <https://doi.org/10.1093/acrefore/9780199389414.013.450>
- Hausman, J. (2012). Contingent valuation: From dubious to hopeless. *Journal of Economic Perspectives*, 26(4), 43–56. <https://doi.org/10.1257/jep.26.4.43>
- Heckman, J. R., Boehmer, E. L., Peters, E. H., Davaloo, M., & Kurup, N. G. (2015). A pricing model for data markets. *iConference 2015 Proceedings*. <http://hdl.handle.net/2142/73449>
- Heckman, J. R., Boehmer, E. L., Peters, E. H., Davaloo, M., & Kurup, N. G. (2015). A pricing model for data markets. *iConference 2015 Proceedings*. <https://www.ideals.illinois.edu/handle/2142/73449>
- Internet of Water. (2018). Methods for valuing data. <https://internetofwater.org/wp-content/uploads/2018/12/01-Options-to-Value-Data.pdf>
- Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., & Spanos, C. J. (2019). Towards efficient data valuation based on the shapley value. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, PMLR, 89, 1167–1176.
- Ker, D., & Mazzini, E. (2020). Perspectives on the value of data and data flows. *OECD Digital Economy Papers*, [No. 299]. OECD Publishing. <https://doi.org/10.1787/a2216bc1-en>
- Koutroumpis, P., Leiponen, A., & Thomas, L. D. W. (2020). Markets for data. *Industrial and Corporate Change*, 29(3), 645–660. <https://doi.org/10.1093/icc/dtaa002>
- Laney, D. B. (2017). *Infonomics: How to monetize, manage, and measure data as an asset for competitive advantage*. Routledge.
- Li, W. C., & Chi, P. J., (2021). Online platforms' creative "disruption" in organizational capital—the accumulated information of the firm [US Bureau of Economic Analysis working paper].
- Mandel, M. (2017). U.S. app economy update. Progressive Policy Institute. http://www.progressivepolicy.org/wp-content/uploads/2017/05/PPI_USAppEconomy.pdf
- Miller, H. M., Richardson, L., Koontz, S. R., Loomis, J., & Koontz, L. (2013). Users, uses, and value of Landsat satellite imagery—Results from the 2012 survey of users. US Geological Survey Open-File Report, 1269. <https://pubs.usgs.gov/of/2013/1269/pdf/of2013-1269.pdf>
- Myers, S. C. (1977). Determinants of corporate borrowing. *Journal of Financial Economics*, 5(2), 147–175.

- New York Times. (2018). "As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants." 18 December. <https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html>
- 'Ocean Protocol, Technical White Paper. (2022). <https://oceanprotocol.com/technology/roadmap#papers>
- ODI. (2018). Data's value: How and why should we measure it? Open Data Institute. <https://theodi.org/article/datas-value-how-and-why-should-we-measure-it/>
- OECD. (2022). Going digital toolbox note: Measuring the economic value of data. OECD Directorate for science, technology and innovation committee on Digital Economy Policy. OECD. [https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/CDEP/GD\(2021\)2/FINAL&docLanguage=en](https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/CDEP/GD(2021)2/FINAL&docLanguage=en)
- ONS. (2019). National Accounts articles: Impact of Blue Book 2019 changes on gross fixed capital formation and business investment. Office for National Statistics. <https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/articles/nationalaccountsarticles/impactofbluebook2019changesongrossfixedcapitalformationandbusinessinvestment>
- ONS. (2021). Valuing official statistics with conjoint analysis. Office of National Statistics. <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/valuingofficialstatisticswithconjointanalysisapril2021>
- ONS. (2022). Conjoint analysis for valuing official statistics. Office of National Statistics. <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/conjointanalysisforvaluingofficialstatisticsuk2022>
- Pei, J. (2020). A survey on data pricing: from economics to data science. In *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/TKDE.2020.3045927>
- Protocol. (2021). <https://www.protocol.com/bulletins/china-big-data-exchange>
- Slotin, J. (2018). What do we know about the value of data? Global Partnership for Sustainable Development Data. <https://www.data4sdgs.org/blog/what-do-we-know-about-value-data>
- Soloveichik, R. (2022). Capitalizing data: A case study of individual credit files. Economic Statistics Centre of Excellence mimeo. <http://escoe-website.s3.amazonaws.com/wp-content/uploads/2022/10/29122657/Capitalizing-Data-A-Case-Study-of-Personal-Credit-Files.pdf>
- Statistics Canada. (2019). The value of data in Canada: Experimental estimates. <https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm>
- Statistics Netherlands. (2021). Estimating the value of data in the Netherlands. IARIW-ESCoE Conference 'Measuring Intangible Capitals and their Contribution to Growth', RSA House. https://iariw.org/wp-content/uploads/2021/10/bondt_paper.pdf
- UN Economic and Social Council. (2022). Recording of data in the National Accounts, March 2022. https://unece.org/sites/default/files/2022-03/6_Recording%20of%20data%20in%20the%20National%20Accounts.pdf
- United Nations. (n.d.). The System of National Accounts (SNA). <https://unstats.un.org/unsd/nationalaccount/sna.asp>
- Veldkamp, L., & Chung, C. (2022). Data and the aggregate economy. Prepared for the Journal of Economic Literature. Forthcoming 2023.
- Verhulst, S. G. (2018). To turn the open data revolution from idea to reality, we need more evidence. Apolitical. <https://apolitical.co/solution-articles/en/to-turn-the-open-data-revolution-from-idea-to-reality-we-need-more-evidence>
- Verhulst, S. G. (2022). The ethical imperative to identify and address data and intelligence asymmetries. AI & Society. <https://doi-org.ezp.lib.cam.ac.uk/10.1007/s00146-022-01410-5>
- WEF. (2021). Articulating value from data. Available at: https://www3.weforum.org/docs/WEF_Articulating_Value_from_Data_2021.pdf
- Yoon, J., Arik, S., & Pfister, T. (2020). "Data valuation using reinforcement learning". In *Proceedings of the International Conference on Machine Learning*. <http://proceedings.mlr.press/v119/yoon20a/yoon20a.pdf>

How to cite this article: Coyle, D., & Manley, A. (2023). What is the value of data? A review of empirical methods. *Journal of Economic Surveys*, 1–21. <https://doi.org/10.1111/joes.12585>