Bennett Institute for Public Policy Cambridge

ODI

Nuffield Foundation

The value of data

LITERATURE REVIEW

LITERATURE REVIEW

The Value of Data

Policy Implications Report

Accompanying Literature Review

CONTENTS

Published: February 2020
Authors: Julia Wdowin & Dr Stephanie Diepeveen, Bennett Institute for Public Policy, Cambridge
Publication from the Bennett Institute for Public Policy, Cambridge in partnership with the Open Data Institute. **bennettinstitute.cam.ac.uk** / **theodi.org**

Introduction

The rise in interest and research surrounding the nature of data and its value runs in line with the explosion of creation, availability and access to data itself. Answering the question of what is the value of data requires first stepping back and making sense of how value from data is created, its capture for different stakeholders and its distribution. Also, currently, most studies treat data as a homogenous entity to which the above question can be applied. However, it will be impossible to make progress in realisations of the valuation of data without the acknowledgment and further study into different types and uses of data. From this expansive view of what is involved in the valuation of data, this literature review lays out the state of existing research into the value of data. It accompanies the project report, which seeks to enrich and expand how value is considered by contributing to a more granular conceptualisation of data.

Across policy literature, there is a broad consensus that there is value from data to be unlocked to the economy as a whole, but also certainty that there are numerous and considerable challenges to doing so. Equally, there is a shared uncertainty as to what the overall value to data might be with these two dimensions hanging in the balance (HMT, 2018; Duch-Brown et al., 2017). It is estimated that £15bn linked to data usage is not being realised due to three main reasons: narrow thinking; a fear of breaching privacy, security and safety; and a belief that the costs of sharing data outweigh benefits (Snaith, 2018). Studies on these points will be discussed below, but while it is clear that the full potential of data use is not being realised, studies point out the significant gains from use already made of data.

The following review begins by discussing the existing relevant literature, both academic and grey, concerning the theory of creation of data value. Current approaches to the valuation of data include determining taxonomies of data that might allow for better valuation strategies by category. A review of this discussion is followed by an exploration of existing literature on the benefits and barriers to data use and the potential weight of these in the process of data valuation.

The second section of the review explores research on existing methodologies for data valuation. This begins by looking at what the literature has identified as important characteristics contributing to the level of value associated to data. It is followed by a review of methods and tools currently conceptualised to value data, such as the use of data value chains. The section ends by reviewing work that addresses the question of redistribution of the value of data.

Literature widely acknowledges the potential for differences in the valuation of data when it is closed, shared or open data. The third section of the review deals with these groups. The idea of data trusts as a response to various concerns voiced over data use and collection is subsequently discussed as covered by existing research.

The literature review subsequently turns to look at the "economics of data". First, the economic characteristics picked up on in existing studies and research are set out, and second, economic approaches to valuing data are reviewed.

The final section looks at the data valuation process in an applied setting, bringing together existing research on data in the specific context of public health.

The Theory

Categorising data

Valuing data can be complex and highly context-dependent (Short and Todd, 2017). To this end, there is a body of existing academic literature that focuses on the process of data becoming valuable in specific contexts. For example, the process has been most widely considered in health (Fontana et al., 2019) and in advertising (Bourreau et al., 2017), but also in the mobility sector (Transport Systems Catapult, 2017) and more broadly in well-being levels, public services transformation, Artificial Intelligence (AI) and machine learning (HM Treasury, 2018). Distinguishing the different areas becomes important when considering questions of data valuation. Across different firms and sectors, it is necessary to consider factors such as the types of data involved, their origin, the way they are used and institutional context (e.g. within a multinational enterprise or not), in order to approach the task of valuing it (OECD, 2019). One of the first points to note is that a distinction must be observed between data and databases, where data should be seen as singular and separate observations in a raw form, while a database is formed once these observations are put together in a structured way, and is ready to be handled (Statistics Canada, 2019).

Data as an intangible asset

Mandel (2012) has suggested that data is complex as a concept to categorise since it is neither a good or service. Data is intangible, like a service, but can easily be stored and delivered away from place of production, like a good. As a result, statisticians need to be thinking about goods, services, and data, in order for policy-makers to build a more accurate picture of economic growth (Mandel, 2012). The measurement of intangible capital and its inclusion in national accounting systems and economic evaluation has an established and growing academic literature (Corrado et al., 2009). However, big data considered as an intangible asset, and measured as one is still not widely researched within this branch of literature (Savona, 2019). Savona (2019) further formulates the need to address the question of whether “the value generated by these different items can be considered homogenous, at least to the extent that they can be safely included in the same (intangible) capital stock.”

Classification systems

It is notable that a vast number of studies investigate data in different contexts as a homogenous entity, or oftentimes treat “data” as personal data, which is revealed by individuals in the context of their leisure or work activities online (Savona, 2019). Some studies have nonetheless begun to approach the undertaking of data-type classification.

One approach to classifying data in a way that may be useful is proposed in the Statistics Canada (2019) report. This organises the information according to what it is about or represents (e.g. weather data, sports data, economic data), as opposed to the alternatives of classifying information based on the applications or services they provide.

Data can also have a number of sources, providing another categorisation method which could be meaningful for understanding valuation. There is 'authored data', which is created through a human process; 'user provided data', which is purposefully provided into a system without expectations; 'captured data' which is data recorded from events in reality or in software; and 'derived data' which is generated by combining, aggregating or processing data (PWC, 2019).

A PWC report (2019) also distinguishes between five categories of data broadly tied to its informational content: 'master data' which describes people, places, and things that are critical to a firm's operations; 'transactional data' which describes an internal or external event of transaction; 'reference data', or information that is used solely for the purposes of categorising data; 'metadata' which characterises other data simplifying the process of retrieving, interpreting or using data; and 'unstructured data' which lacks a consistent format to describe objects and attributes (PWC, 2019).

Dodds (2019) provides a template series of considerations by which to categorise different types of datasets as follows, enumerating ten different types of dataset that take into account:

- the characteristics of the dataset
- the capabilities of the publishing organisation and the funding they have available
- the purpose behind publishing the data
- the ethical, legal and social contexts in which it will be used.

The ten observed datasets are as follows: The Study, The Statistical Index, The Sensor Feed, The Register, The Database, The Description, The Personal Records, The Social Graph, The Observatory, and The Forecast.

The Swedish National Board of Trade (2014), reporting on the role of cross-border data transfers for companies in Sweden, categorise types of data by usage (those types used by companies), and relate the following: corporate data; end-customer data; human resources data; merchant data; and technical data.

The OECD (2011, quoted in The National Board of Trade, 2014), furthermore, provides taxonomies of data type by the way in which it is generated. These include: volunteered data (created and shared by individuals, for example, on social network profiles); observed data (recordings of the actions of individuals, for example, browser history); and inferred data (data obtained from analysis of volunteered or observed data).

The Statistics Canada (2019) report, mentioned above, further points to looking to a categorisation of asset owners in the context of data, making a distinction between legal owners of assets and economic owners. Legal owners determine who can use the asset and the terms under which it can be used, while economic owners use the given asset and bear any subsequent risk associated with it (Statistics Canada, 2019).

Benefits from data use

Various benefits from data use are considered across the literature, ranging from those that can be gained by data holders directly and indirectly, and those that can be of public benefit. These are found to be linked positively to enhanced innovation, social welfare and public service delivery.

The Government Office for Science (2018) lists broadly a number of benefits to data holders. These include reputation gains, benefits to humanitarian causes, reduced operational costs or in some cases direct revenue returns.

The added-value to firms and other entities of the collection and use of data is widely reported (Snaith, 2018; Higson and Waltho, 2009; McKinsey, 2013; LaValle et al., 2011; Deloitte, 2019). One study (LaValle et al., 2011), based on the findings from surveys of three thousand business executives from organisations and academic experts located around the world, concludes that the best performing organisations are twice as likely to apply analytics ('using data to enhance business performance') (Deloitte, 2019) to activities. PWC finds that valuations of data-driven firms within the same industry tend to be higher than those of their peers, and furthermore, that companies with data analytics capabilities are twice as likely to end up in the top quartile of performance within their industries (PWC, 2019). Firms stand to gain in productivity from the use of relevant data in their decision-making processes; HM Treasury (2018) find that data-driven decisions can result in 5-6% higher output and productivity. In one concrete case of production system digitalisation, BP increased production by 30,000 barrels a day, through data use (Raval, 2019). In another case, similarly, intermediaries, such as credit reference agencies, attributed £23m/year of their revenues and £5m/year of their costs to their use of Companies House data, which they used as input to their own data products and services.

The McKinsey Global Institute and McKinsey Center for Government (2013) have focused on the trend of open data, and released a report on the potential of data in "unlocking innovation and performance." The report enumerates benefits to making data more liquid, noting improving the efficiency and effectiveness of existing processes; making possible new products, services, and markets; and creating value for individual consumers and citizens (McKinsey, 2013). The report also (McKinsey, 2013) suggests consumers stand to gain by saving money through greater price transparency and using more information to make decisions.

Studies find that there are social and public flows of benefits from data use in certain contexts, too. Digital Science (2019) point to the value created by open public data in terms of institutional transparency and accountability, and the potential it has to encourage innovation.

It must be noted that the value added to firms from the availability of data and the use of this data may differ along the lines of what type of firm we are dealing with. Bergemann and Bonatti (2018) discuss different pricing strategies of information markets depending on whether we are treating competing firms, monopolists or oligopolists, for example. In addition, the sector context is important as Deloitte (2019) note in a published report that, "investment in analytics in the SME sector is set to be a key differentiator."

Barriers to data use

As mentioned, concerns arise across existing literature regarding the possibility that data use is not being fully realised. Research finds a number of barriers to data use including potential costs to the private and public sectors, and individuals, as well as risks associated with its use. Snaith (2018) goes as far as to conclude that the costs of sharing data outweigh benefits.

Data as liability

Data is found to be a potential liability given data protection law requirements and the legal uncertainty surrounding its management that can arise (Higson and Waltho, 2009; HM Treasury, 2018; Spiekermann et al., 2015). Overall, especially regarding big datasets, there are significant challenges to "storing, protecting, accessing, and analyzing massive amounts of data" to organisations (Short and Todd, 2017). Datasets with personal data, as well as others, can become targets of cybercrime, especially in the case of financial information (Spiekermann et al., 2015). There are also costs and resource requirements involved in data use (Government Office for Science, 2018), including the infrastructure needed for data analytics.

Despite the gains available to firms, from their survey-based research, Short and Todd (2017) found that the majority of companies "had no formal data valuation policies in place." The effort involved was repeatedly said to be time-consuming and complex. The McKinsey Open Data report (2013) backs this finding, remarking that there are costs in the "effort to measure, analyze, and incorporate insights from the data into daily decisions for both consumers and businesses."

Privacy concerns

The McKinsey Open Data report, as well as Bergemann and Bonatti's paper (2018), acknowledges data-related privacy concerns, and the potential costs of the loss of freedom with the increase in open data availability. On this, Fuller (2019) states that "it is costly to measure product attributes (such as "privacy")."

Tonetti and Jones (2018) find that at the stable equilibrium of demand and supply of data, firms may not effectively respect the privacy of consumers. Acquisti et al. (2016) have concluded that in online settings, consumers are not in a position to make informed decisions about the privacy of their personal data. It is the risk of harm from the potential to gain insight into areas sensitive to individuals from inappropriately shared data that motivates data protection laws (European Commission, 2019).

For firms, there may be concerns related to the need to protect competitive advantage (Government Office for Science, 2018).

Data valuation: methods and dimensions

The categorisation of data, treated in studies discussed above, is important for the subsequent valuation of data, because, as Savona (2019) argues, it is evident that different types of data are subject to the data value chain, and only some of them may be valuable.

Valuation characteristics

A number of studies have addressed the task of valuing data. In most of these cases, the estimation of the value of data has been directly linked to market valuations or value of transactions. The scope of factors involved in the valuation of data are considered by Short and Todd (2017), who enumerate the following as components on which value will be based: usage type and frequency, content, age, author, history, reputation, creation cost, revenue potential, security requirements and legal importance. They note also that data value may change over time in response to new priorities, litigation or regulations. PWC (2019) reports that the value of data is accrued through the following characteristics: “exclusivity, timeliness, accuracy, completeness, consistency, use restrictions, interoperability, liabilities and risk.” Snaith (2018) points out potential differences in data value due to the different desired outcomes, economic systems and social contracts of governments and businesses. Moreover, Belleflamme (2018) attributes an increase in the value of data from a demand-side perspective to the “increases in their volume, variety, veracity and velocity.”

If data can serve a specific purpose or contribute to understanding a particular cause, for the given entity it can take on greater value. LaValle et al. (2011) discuss how data takes on greater value if organisations can first identify their questions and needs. In turn, relevant data acquires greater value to the organisation. An OECD report similarly maintains that, ultimately, the value of data to businesses will depend on how and where in the business value chain data is put to use (OECD, 2019). Another important element in determining the value of data, and a “fair valuation of data” involves anticipating and accounting for future uses of data (Spiekermann et al., 2015).

There is the question of timeliness in gains in value of data (LaValle et al., 2011; PWC, 2019), where, if inopportune, data can be without value to the entity or lose value. HM Treasury (2018) point out that the volume of data does not reflect its value; if data is inaccessible or irrelevant, however large the amount there is, its value will remain unrealised.

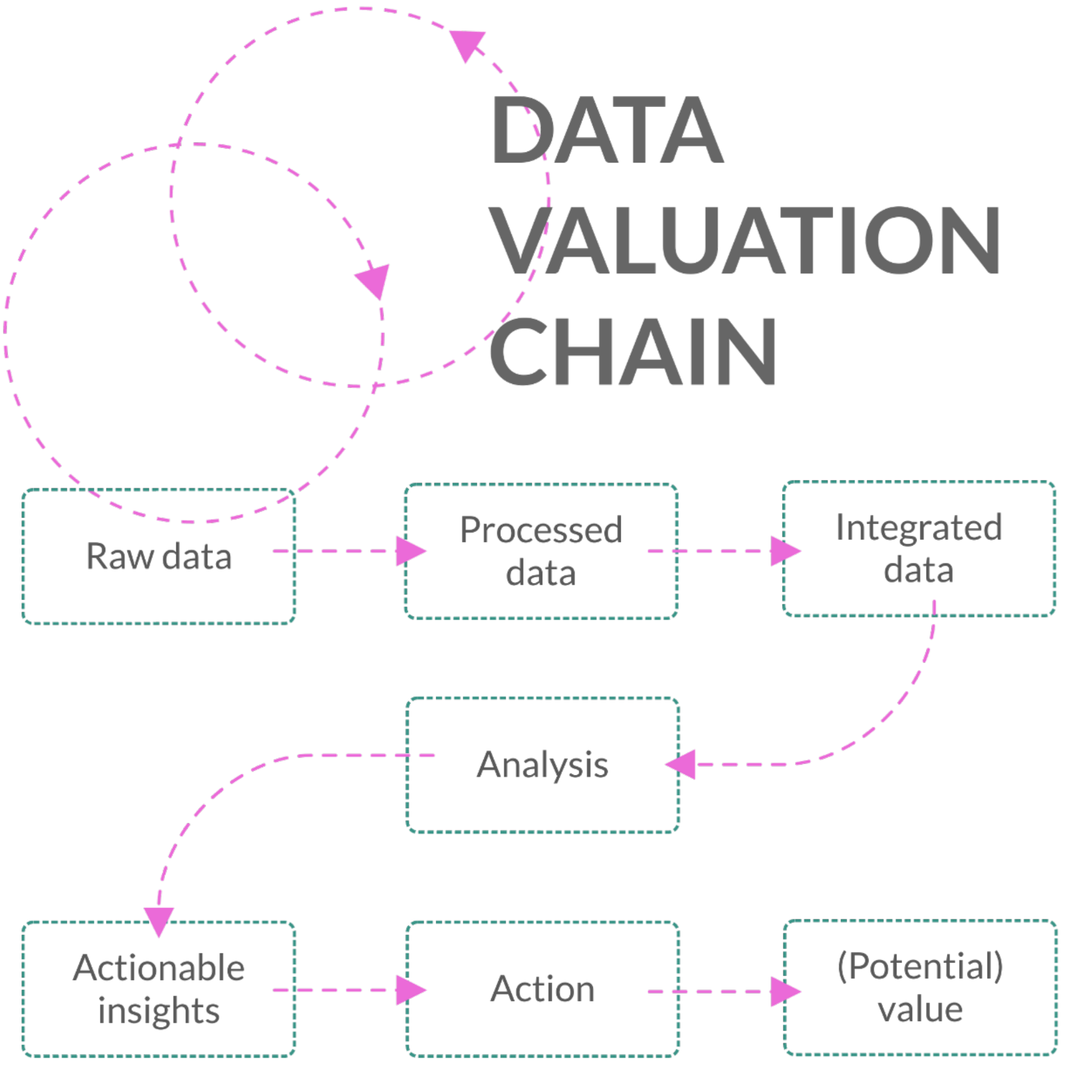
PWC (2019) add that for a firm that has acquired data, the less restricted the use of the data is within the firm, the higher its value to that firm. In addition, the more unique the dataset, the more valuable the data is (PWC, 2019).

Valuation methods

Among the three possible data valuation methodologies (income, market and cost approach), the PWC report adds that in cases where data can be reproduced or replaced, “the Cost Approach can provide useful upper and lower bounds for valuation” (PWC, 2019). In terms of evaluating the value of data to firms, Mawer (2015) suggests an income-based valuation approach, where “value is defined based on an estimate of future cash flows to be derived from the asset”, with a view to data as an intangible asset. The valuation processes indicate that determining the value of data is subject to the analytics process the given data goes through. Savona (2019) explains that data value is obtained by data capitalists “through investments in digital infrastructures, organisational and human capital, enabling data collection and aggregation, treatment and analysis.”

The Global Partnership for Sustainable Development, in a report on the value of data to SDG measurement, reviewed five different data valuation methodologies: cost-based (value is based on the cost to produce data); market-based (value is based on the market price of equivalent products); income-based (value is based on future cash flow estimations); benefit monetisation (value is based on estimation of monetisation of benefits associated to the data product); and impact-based approaches (value is based on the assessment of the causal effect of data availability on outcomes [social and economic]) (Slotin, 2018). The paper concluded that none of the above methods for data valuation were sufficiently convincing to influence policymakers, but that the impact-based approaches seemed to be the most encouraging for their purpose in evaluating data investments and outcomes that affect lives. After reviewing each of these approaches, the paper concluded, first, that measuring the value of data is very difficult and there is no consensus on how best to do it, and second, that none of the methods are sufficient to influence policymakers.

Data Valuation Chains



Source: Mawer, 2015

(For orignal design, see references)

Regarding the above data valuation chain, Mawer (2015) begins by determining that data's value increases as it moves through the data valuation chain, and valuation depends where on the chain data lies. Similarly, Corrado (2019) outlines a value chain in which raw individual data carries the lowest value, and that the aggregation of data, and its treatment, add or give it value (Savona, 2019).

Secondly, Mawer (2015) introduces the element of risk involved in dealing with data – it may turn out not to have value (see also Savona, 2019). Additionally, the view that the valuation chain, depicted above, must be fully completed to realise (potential) value from data introduces a further complication in the quantification of data's value (Mawer, 2015).

LaValle et al. (2011) point out, though, that the nature of the data available does not need to be complete in order to take on value. They conclude that "contrary to common assumptions, it [the value of data and its analytics to a firm's productivity gain] doesn't require the presence of perfect data or a full- scale organisational transformation."

PWC's report (2019) puts forward four main considerations for an organisation in making sense of the value of the data they may hold by "performing an inventory of its data." It does this by (1) assessing the quality of its data, (2) identifying gaps in the data, (3) highlighting key issues which might hinder its data strategy (such as legal or regulatory restrictions over use of the data), and (4) thinking about possible applications for the data.

Technological innovations may also determine new ways in which to think about or approach valuing data. Arrieta Ibarra et al. (2017) comment on the growing and future importance of machine learning for data valuation, creating new potential uses and services and, therefore, new future income streams. Machine learning can allow for the estimation of the marginal effect of new data on predictions. This is a novel method for data valuation, and one being pursued, for example, by Microsoft (Koh and Liang, 2017, ref. Arrieta Ibarra et al., 2017).

Moreover, there is some discussion on what does not equate data with value. For example, Short and Todd (2017) note, "IT functions alone cannot make the decisions that transform data into business value." There is separation of the existence of available data from its instant transformation into value.

There is consensus across studies that data takes on value only in the context of its combination with other data. The OECD (2019) states that personal data collected only takes on value when it is aggregated with other data; on its own, it may have close to no value. They write, "once those data points [that are collected from the users of online social media platforms] are transferred and aggregated with millions of other data from across the globe they become the basis for data analytics and thus for value creation."

In sum, this review of the existing literature indicates that although there is discussion of approaches to data valuation, these are largely market-based valuations, and there is scarce discussion of approaches to the estimation of the social or public valuation of data.

Personal preferences in data valuation

Concerns about privacy issues have been raised in studies discussed above, making data a liability. However, from a perspective of consumer preference for privacy, Winegar and Sunstein (2019) note that it is plausible that, for some part of society, "having personal data collected and the associated results, for example, receiving ultra- targeted ads, experiencing greater ease of purchasing products online, and so forth, provide a positive value, potentially greater than the privacy concerns associated with sharing their data." The valuation of privacy, then, is not a straightforward issue. This perspective supposes this brings an aspect of greater complexity for data valuation itself. To this end, Spiekermann et al. (2015) find that identifying individual privacy preferences and valuations, and, subsequently, determining its effect on data valuation, is often a difficult and inconsistent activity. This is due to heuristics and psychological biases which affect preferences and, as a consequence, individual valuations of personal data (Acquisti 2004; Acquisti et al. 2013 via Spiekermann et al., 2015).

The redistribution of the value of data

As the availability of data and data access expand rapidly, literature on the value of data has raised the pertinence of interrogating the redistribution of data. Some studies have gone as far as suggesting approaches to framing this problem.

Savona (2019) discusses three possible frameworks for redistributing the value of data. The first is to consider data as capital, whereby data collection and their analytics are considered an investment in intangible assets. Practically, she suggests this would involve supranational public institutions creating a data market in order to be most effective. The second framework involves viewing data as labour; this would include compensating individuals for their data, and imposing a corporate tax on digital activities (see Ibarra et al., 2018; Posner and Weyl, 2018). The third approach relies on seeing data as Intellectual Property, which as a usable and licensable asset elicits financial reward, and the protection of intellectual property rights.

Within a framework that considers data as labour, the production of data should be viewed as labour. Arrieta Ibarra et al. (2017) discuss the consequences of this. Developing this framework, Snaith (2018) concludes that data labourers could organise a data labour union to certify data quality and ensure that labourers reach potential earnings.

Bergemann and Bonatti (2018) identify an increasingly urgent need for compensation for data, maintaining that with growing awareness about data sharing practice, producers of data will need to be compensated in order for them to continue to be willing to reveal information. A distinction, as relayed in the recent Statistics Canada report (2019), therefore needs to be made between data that is produced by businesses and governments for their own use but not sold in the marketplace, and data that is supplied by households to businesses and governments "as payment-in-kind in exchange for other services, as for example in the case of Facebook, Google and many other online services." The report also goes on to emphasise that since businesses and households spend significant resources protecting data, we can observe that data is, in some way, 'owned'. This leads to the conclusion that data is produced. Spiekermann et al. (2015) observe highly limited freedom of "producers" of data in this context, noting that market forces "render people into "data subjects" whose "digital identities" are traded and used (potentially without their knowledge and consent).

The World Bank (Connon and Djankov, 2019) has recently raised the question of how the benefits reaped from data collection and its use should be distributed. They ask, in the public sector, is it fair that governments keep all the revenues and citizens, from whom the data was collected, receive nothing? They suggest this as the reason for which the more recent proposition of digital taxes is gaining traction, as a way of re-distributing profits from data usage more fairly.

Closed, shared and open data

Data can be conceptualised as being located on a data spectrum that shifts from closed to shared to open data. Individuals' data can be found across this spectrum, and its position determines the access to it, and, in turn, its value (Open Data Institute, 2019).

Open Data

Open data is understood as information that can be freely used, modified and shared by anyone for any purpose and must be available under an open licence (European Commission, 2015). The European Data Portal (2017) stresses the potential impact of open data: "when opened, data can become a force of growth and development for all countries, regardless of geography and level of economic development."

Economic benefits linked to increased open data availability have been enumerated across literature in this area. For companies, benefits have included enhanced innovation, reduced costs and increased efficiency (European Data Portal, 2017). The economy has also seen an increase in the demand for Open Data workers in response to the expanding market (European Data Portal, 2017; European Commission, 2015). The HM Treasury (2019) expect that open or shared data promotes competition and innovation in the economy.

Open data has significant potential to enhance, perhaps especially, public services by providing tools to address a wide range of societal challenges: growth of the knowledge economy; encouraging the participation of citizens more actively in political and social life; saving costs and saving time for users of Open Data applications; increasing transparency and accountability; and increased efficiency by sharing data between public administrations (European Data Portal, 2017; European Commission, 2015; Open Knowledge Foundation, 2019). The European Data Portal provides estimates for the value of open data as a percentage of GDP ranging from 0.08 percent to 7.19 percent (European Data Portal, 2017). The OECD estimates this range to be from 1 percent to 2.5 percent of GDP (OECD, 2019). The Government Office for Science further provides another estimate; they estimate the benefit to the open nature of public data to sit between 0.5 to 4 percent of GDP. In transport, they suggest that Transport for London has been realising between £15m to £58m of time savings for passengers due to data sharing." In yet another estimate, the Open Data Institute concluded that certain public datasets being made open data would create 0.5 percent more GDP growth per year for the British economy as opposed to a case where users pay for access to the data (Open Data Institute, 2016).

Thus far, the Open Knowledge Foundation (2019) suggests that open public sector data that has been found to have the most value are: geospatial, earth observation and environment, meteorological, statistics, companies and transport.

The European Commission, however, reminds that data alone does not provide economic value, but rather the insights that can be gained from it. This must be reiterated in the case of Open Data, and the benefits that can be gained from its analysis. In addition, they note that in a large number of instances Open Data is combined with other types of data to get greater value from it, for example, Big Data, private data etc. This increases the potential value of each data source (European Commission, 2015).

The value of Open Data in particular may depend on the sophistication and availability of adequate infrastructure that supports and allows the analytics of this data, thus depending on the quality of insights to be gained from the data. The European Commission (2015) comments that we do not have at our disposition a pan-European data portal infrastructure, aggregating metadata records of public data resources from all over Europe, which is a "key requirement for a healthy Open Data ecosystem."

Costs associated with Open Data initiatives include transformation and adaptation costs, infrastructure costs and structural management costs (identified by Technical University Delft via European Data Portal, 2017).

Sharing data

Literature has also addressed the particular benefits to be gained from sharing private data. One of these is the direct benefit to better policy-making. Larger and better quality data collections allow the public sector to design more focused evidence-based policies (Verhulst, 2019). Verluhst (2019) also writes that data partnerships and exchanges can be helpful for more accurate predictive functions in decision-making.

The Government Office for Science (2018), however, points out that, despite the evident benefits, the sharing of private data will not likely occur spontaneously without incentives to the holder of the data. The Government Office classifies three private data sharing models: 1) Data pools, where datasets from different sources are pooled together and kept in one organisation; 2) Data sharing platforms, where access to multiple datasets is available on a platform; and 3) Direct data sharing, where a single organisation makes its data available for re-use.

The type of data, the benefits available and the incentives to the data holder are factors that will all influence the success of government interventions to promote the sharing of private data and its regulation (Government Office for Science, 2018).

Following on from the above discussion about data's potential role in public policy formulation, there is some consideration in recent literature on the links between data sharing, governance and trust. Pope (2019) questions how the more extensive use of data for delivering services will impact upon citizens' "mental models of government", and asks if they will trust it. He cites studies that indicate risks to data collection that arise from fears that data collected for one purpose will be used for other purposes (Pope, 2019). He concludes that data infrastructure must be cultivated in ways that bear in mind public good and accountability measures.

Data Trusts

The idea of data trusts is found across numerous papers and reports, and proposed as a solution to (1) a number of the concerns raised related to data, and (2) a way of distributing the value of data from users to contributors. Mainly, they are perceived as a potential way to compensate for missing data markets (London Economics, 2019), and to increase transparency and accountability in data handling, thereby addressing privacy concerns.

Delacroix and Lawrence (forthcoming) assert that data trusts have a number of models: legal trust, contractual, corporate, public, and community trust models. The ODI (Hardinges, 2018) have also found different interpretations of 'data trusts', which include:

- A data trust as a repeatable framework of terms and mechanisms.
- A data trust as a mutual organisation.
- A data trust as a legal structure.
- A data trust as a store of data.
- A data trust as public oversight of data access.

There is often language related to safety used in the context of the purpose of data trusts. Snaith (2018) describes the growing call for data trusts that would work to "ensure that data exchanges are 'secure and mutually beneficial' for all stakeholders." Additionally, the Hall and Pesenti Report (2017), which discusses the implementation of data trusts, recommends data trusts as a means to "share data in a fair, safe and equitable way."

Some studies highlight ways in which data trusts could attend to other public concerns, especially those regarding privacy and data governance. Croft (2019) outlines in a piece in the Financial Times how data trusts may offer individuals the opportunity to become trustees and in this way be part of the governance of their anonymised data. Views on the potential benefits of data trust are also raised in academic and grey literature. Data trusts can increase the confidence and willingness of individuals and firms with regards to data handling. There could be an opportunity to establish certain conditions for the quality of the data provided by members (Pinsent Masons et al., 2019). Going back to Croft (2019), he suggests the potential for trusts to encourage companies to collaborate on projects with a common goal by sharing data, for example, supermarkets targeting food waste.

There are also various problems connected to data trusts that have been signalled. The regulatory framework of such trusts is not clear, and it could increase the risk of hackers obtaining data, which would disincentivise governments, firms or individuals to share their data in the trust (Croft, 2019).

In terms of regulation in the context of data trusts, it has been suggested that a Data Trust law could exist alongside the GDPR (Delacroix and Lawrence, 2018, quoted by Savona, 2019). In this way, data trusts could circumnavigate the issues around 'data property' by focusing on 'data rights' instead.

The Economics

The Economic Characteristics of Data

A report by Open Data Science labels data as an asset, which as such grows in value through use (Open Data Science, 2019). Data can accumulate but unlike other produced assets does not physically decay or deplete naturally over time, although its relevance, timeliness, and a whole number of other pertinent factors mean the economic value of data may depreciate over time (Statistics Canada, 2019; Savona, 2019).

The idea that data is the new oil is contested throughout current literature (Snaith, 2018; Jones and Tonetti, 2018). Data has, and can be used at, a zero marginal cost (while data infrastructure and analytics do assume costs) (OECD, 2019). Also, the same data can exist simultaneously in multiple places (Statistics Canada, 2019), which leads it to be considered as non-rivalrous in nature (Jones and Tonetti, 2018). Savona (2019) suggests that personal data is both non-rivalrous and excludable, and that the combination of these two characteristics means data can be considered club goods.

The non-rival nature of data gives rise to a number of questions related to property rights and ownership of data (Snaith, 2018; Duch-Brown et al., 2017; HM Treasury, 2018), and further discussion about the consequences of non-rivalry for establishing the rights to use, exclude and transfer data (HM Treasury, 2018). The non-rival nature of data can lead firms to choose to hoard data they own (Jones and Tonetti, 2018). From this perspective, this characteristic of data, then, could inherently become a barrier to sharing data.

Looking beyond the characteristics of data itself, data analytics can produce knowledge. Knowledge then becomes the entity subject to valuation, and becomes a public good (Savona, 2019). Social benefits can also be derived from the knowledge as public good. However, when it becomes property of private entities the knowledge becomes a private good (Savona, 2019), and may result in different valuation.

In economic terms, there is often an asymmetry of information around the handling and treatment of data. Asymmetry of information often emerges as a barrier to data sharing in the form of a coordination problem. Uncertainty can be seen in the dynamic between producers and consumers of data. The incentives for data producers to produce or share their data is not always known, while data consumers often do not have full information regarding the quality, amount or type of data that is available (London Economics, 2019). Winegar and Sunstein point out that consumers typically have “highly imperfect information about whether their data is collected, which data is collected, and how their data is used by online advertisers” (Winegar and Sunstein, 2019).

Some studies also refer to the externalities that occur from data production or data use, where the positive and negative externalities are subsequently linked to the under- and over-production of data respectively (see Bergemann and Bonatti, 2019). However, research on how these externalities in relation to data may be valued is sparse. Bergemann and Bonatti (2019) highlight the additional value from informational content to be gained through the aggregation of individual data. When individuals' personal data is combined with observations of other individuals, its individual marginal value of zero in isolation takes on a value (Statistics Canada, 2019). Still, externalities, and potential additional value emerging from the combination of datasets, are not widely covered in the literature.

Economic valuation

Some economic literature looks at the economics of data availability and sharing, that is, the mechanisms of demand and supply to understand how value of data is created. From a consumer or demand perspective, the nature of the information collected, and its potential or actual uses, can determine, firstly, value (Belleflamme, 2018) and secondly, a consumer's willingness to share it (Bergemann and Bonatti, 2018). From the supply-side, questions that address data valuation include where the data comes from, what their quality is, and who controls production and collection (Belleflamme, 2018).

Applied Data Use

Public health and data

A growing literature on the role and value of data in the context of health, mainly public health, appears to form a consensus about the specific benefits available to the development and progress of public health as a result of data use and analysis. Equally, this literature highlights a range of challenges posed by data use that are particular to the field of health. The demand and interest in public health data is on the rise alongside a rise in acceptance of evidence-based medicine. It is argued big data algorithms that provide the best source for evidence (Groves et al., 2016). A McKinsey report (Groves et al., 2016) identified more than 200 businesses that had been created since 2010 that were developing a diverse set of innovative tools to handle available healthcare data.

Studies have also found that unusually large and full data sets are created and available in the context of public health. For example, Fontana et al. (2019) note that NHS data can be considered a unique country-wide longitudinal dataset. Indeed, they argue it is one of the most comprehensive, longitudinal patient data sets in the world.

Looking to questions about value creation, data use in health can lead to discoveries, solutions and benefits to patients, in some cases life-saving. Also, for healthcare professionals, data use can help to lower costs and improve care (Fontana et al., 2019; Raghupathi and Raghupathi, 2014). Specifically, data has the potential to help identify the most clinically and cost effective treatments. It can also inform analysis and tools that are used to various ends, including: to proactively identify individuals who would benefit from preventative care or lifestyle changes; identify predictive events and support prevention initiatives; assist patients in determining the care protocols or regimens that offer the best value; identify, predict and minimize fraud by implementing advanced analytic systems for fraud detection; and implement much nearer to real-time, claim authorization (Raghupathi and Raghupathi, 2014). When considering potential benefits of data use, Groves et al. (2016) approach the value of healthcare data based on a concept of value derived from the balance of healthcare spend and patient impact.

The value of data may be linked to the value that is placed on potential returns from this data use and analytics. Further to this, Fontana et al. (2019) point out that the potential for returns increases the likelihood of firms, industries requiring exclusivity in respect of its use of data “to preserve its competitive position.”

Still, for the above benefits to be realised, the quality of available data and the sophistication of the associated analytics must be taken into account. Raghupathi and Raghupathi (2014) note that “data quality issues are of acute concern in healthcare for two reasons: life or death decisions depend on having the accurate information, and the quality of healthcare data, especially unstructured data, is highly variable and all too often incorrect.”

In a similar vein to discussions of the risks related to healthcare data quality, there is widespread attention and acknowledge in a large part of the literature on data in healthcare on challenges and risks in using data in this specific context. It is widely noted that this is a highly sensitive environment in which to be handling data, and, indeed, in many cases it involves sensitive data. Issues that require particular discretion include privacy (Groves et al., 2016; Fontana et al., 2019), security, distribution, re-use (Fontana et al., 2019), 'ownership' of data, and the regulatory load associated (Feldman et al., 2012). A case study of DeepMind-Royal by Powles and Hodson (2017), concludes that there is much at stake with regards privacy and security with artificial intelligence structures such as Google DeepMind "being given unfettered, unexamined access to population-wide health datasets". They suggest that as a consequence, structures like Google Deepmind will "build, own and control networks of knowledge about disease." This is because, the authors conclude, it cannot be known what Google and DeepMind "are really doing with NHS patient data, nor the extent of Royal Free's meaningful control over what Google and DeepMind are doing" (Powles and Hodson, 2017). Powles and Hodson (2017) further argue for a "meaningful agency" that patients must have with regards to data collected on their subject. They take the view that the "value embodied in these NHS datasets does not belong exclusively to the clinicians and specialists who have made deals with DeepMind", but it also belongs to those who generated it.

Another risk, perhaps especially acute in healthcare, is tied to computational errors within data use. These can be particularly costly in this context (Powles and Hodson, 2017).

These enumerated issues indicate a growing tension arising from the question of what is appropriate or right for a patient, and what is right for the healthcare ecosystem (Groves et al., 2016). In the case of the study of DeepMind – Royal Free, which presents an example of the use of machine learning and artificial intelligence in healthcare, researchers determined, in terms of patient autonomy and public value, that "existing institutional and regulatory responses are insufficiently robust and agile to properly respond to the challenges presented by data politics and the rise of algorithmic tools in healthcare" (Powles and Hodson, 2017).

The presence of restrictions and safeguards around the use of healthcare data also implicate value creation. Safeguards that are already in place can mean that, despite the quantity and richness of data available, data is not used as optimally as could be possible. For example, Groves et al. (2016) argue that as a consequence of safeguarding it may be more difficult to integrate different sets of patient data. They find that "important information often remains siloed within one group or department because organisations lack procedures for integrating data and communicating findings". Feldman et al. (2012) point out that due to the high volume of data available in healthcare, a critical issue is discovering better ways to handle existing data, thereby exploiting "the disconnected puddles and lakes of existing data (e.g., health records, clinical trial data, actuarial information)."

Fontana et al. (2019) find that in the valuation of public health data, factors that contribute to the creation of value include objective dimensions such as accuracy, integrity, consistency, completeness, accessibility, precision, timeliness, linkage of the data at hand and subjective dimensions such as relevance, usability, believability, clarity, objectivity, scarcity. Specifically, Fontana et al. (2019) identify that the approach to valuing data in a public health context will depend on "the quality of the data, the type of end product, the extent of its reliance on NHS data, the availability of other appropriate datasets, the extent of the NHS's inventive contribution, and how much work and costs the commercial partner will need to invest to make the data useful."

On this point, a few authors note that the volume, veracity, velocity and variety of data in the health context has expanded dramatically. To this end, the analytic techniques needed to acquire, use and therefore attach value to data have grown in complexity (Raghupathi and Raghupathi, 2014). These four characteristics of big data - volume, veracity, velocity and variety - still hold in the context of healthcare data, though Feldman et al. (2012) note that "volume" is particularly outstanding in healthcare. They state, "the enormous variety of data—structured, unstructured and semi-structured—is a dimension that makes healthcare data challenging." There is perhaps a trade-off increasingly at work between quantity and quality, or, in big data terms, between volume and veracity: the increased variety and velocity of healthcare data availability, plus the sheer quantity, make the process of cleansing and analysing data somewhat more complex, but high-quality data is needed and desired to work towards advances and improvements in care, costs and treatments.

Different factors have also been found to support value creation in health. Looking to data sharing and communication, some literature finds a potential for partnerships to facilitate the exchange of data to the benefit of entities involved as well as to patients (Fontana et al., 2019). The McKinsey report notes the potential for value capture that can be created through partnerships. Players in the healthcare setting "are more likely to capture value from big data by developing innovative partnerships and aligning goals with organisations that have traditionally been their competitors" (Groves et al., 2016).

Concluding Remarks

The above literature indicates a growing and perhaps increasingly urgent search to understand how data takes on value, the nature of the data that takes on value and why. However, as things stand, there are more questions than answers, and there is a need for further research down a number of pathways. There is a need to better and more systematically understand the categorisation of different types of data, and how data might be categorised in ways that help to explain the processes (potentially different) through which these different types of data take on value. There is also limited literature on the process of value created by the volume of data versus the content of data, or the gains and positive or negative externalities to be found from the interaction of combining datasets, where the value of data cannot be equated to the sum of its parts, a question that is nonetheless raised in a number of the papers reviewed. Finally, existing literature indicates to an opportunity to further explore the potential consequences of specific contexts on the value of data: including exchanging versus sharing data; the degree and potential substitutability of datasets in different contexts; and the differences between ex ante (information structures) and ex post (concrete data realisations) data sales (Bergemann and Bonatti, 2018).

References

Acemoglu, et al., (2019), 'Too Much Data', MIT Working Paper, September 2019, https://economics.mit.edu/files/17760.

Arrieta Ibarra, I., Goff, L., Jimenez-Hernandez, D., Lanier, J. and Weyl, G., (2018) "Should we treat data as labor? Moving beyond "free"." *AEA Papers and Proceedings*, 108: 38-42.

BEIS and Companies House, (2019), "Companies House data: valuing the user benefits", available at: https://www.gov.uk/government/publications/companies-house-data-valuing-the-user-benefits.

Belleflamme, P., (2018) "Economic Models of Data", Working Paper, August 2018.

Bergemann, D., and Bonatti, A., (2018), "Markets for Information: A primer." CEPR Discussion Paper 13148.

Bourreau, M., Cailaud, B., De Nijs, R., (2017) "The Value of Consumer Data in Online Advertising", Paris School of Economics Working Paper, 2017.

Brynjolfsson, E., et al., (2019), 'Using massive online choice experiments to measure changes in well-being', *PNAS*, 116 (15), pp. 7250-7255.

Connon, D., and Djankov, S., (2019), "Valuing user data", World Bank blogs, available at: https://blogs.worldbank.org/developmenttalk/valuing-user-data.

Croft, J., (2019), "Data trusts raise questions on privacy and governance", *Financial Times*, 13 September 2019, available at: https://www.ft.com/content/a683b8e4-a3ef-11e9-a282-2df48f366f7d.

Dekkers, M., et al., (2006), "Measuring European Public Sector Information Resources Final Report of Study on Exploitation of public sector information – benchmarking of EU framework conditions".

Delacroix, S., and Lawrence, N., (2018), "Bottom up Data Trusts: Disturbing the 'One Size Fits All' Approach to Data Governance", *International Data Privacy Law*, Doi.org/10.1093/idpl/ipz014.

Deloitte, (2019), "Generating value from data capture", available at: https://www2.deloitte.com/ie/en/pages/deloitte-private/articles/generating-value-data-capture.html, accessed on 09.10.2019.

Dodds, L., (2019), "How can we describe different types of dataset? Ten dataset archetypes", Blogpost available at: https://blog.ldodds.com/2019/05/30/how-can-we-describe-different-types-of-dataset-ten-dataset-archetypes/, accessed on 09.10.2019.

Duch-Brown, N., Martens, B., Mueller-Langer, F., (2017), "The economics of ownership, access and trade in digital data", European Commission JRC Digital Economy Working Paper 2017-01.

European Commission, (2015), "Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources", Digital Agenda for Europe, 2015, available at: https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.

European Commission, (2019) "Free flow of non-personal data", available at: https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data.

European Data Portal, (2017), "The Economic Benefits of Open Data", Analytical Report 9, available at: https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf.

Feldman, B., Martin, E., and Skotnes, T., (2012), "Big data in healthcare hype and hope", *Dr. Bonnie* 360 (2012): 122-125.

Fontana, G., et al., (2019), "Ensuring that the NHS realises fair financial value from its data", *manuscript, forthcoming*.

Fuller, C., (2019), "Is the market for digital privacy a failure?", *Public Choice*, 180: 353–381.

Groves, P., et al., (2016) "The 'big data' revolution in healthcare: Accelerating value and innovation", McKinsey, Center for US Health System Reform, Business Technology Office, (2016).

Hardinges, J., (2018), “What is a data trust?”, Open Data Institute blog, available at: https://theodi.org/article/what-is-a-data-trust/.

Higson, C., and Waltho, D., (2009), “Valuing Information as an Asset”, EURIM presented paper, November 2009, available at: http://faculty.london.edu/chigson/research/InformationAsset.pdf.

HM Treasury, (2018), The Economic Value of Data: A discussion paper.

HM Treasury (2019), ‘Unlocking digital competition Report of the Digital Competition Expert Panel.’ Report available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf.

Jones, C., and Tonetti, C., (2018) “Nonrivalry and the Economics of Data,” 2018 Meeting Papers, 477, Society for Economic Dynamics. Available at: [http://christophertonetti.com/files/papers/JonesTonetti_DataNonrivalry.pdf]

Jin, Y., and Vasserman, S., (2019), ‘Buying Data from Consumers’, NBER Working Paper 2019.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., and Kruschwitz, N., (2011), “Big Data, Analytics and the Path from Insights to Value”, *MIT Sloan Management Review*, Winter 2011, Vol 52 (2).

London Economics, (2019) “Independent assessment of the Open Data Institute’s work on data trusts and on the concept of data trusts: Report to the Open Data Institute”, available at: https://theodi.org/wp-content/uploads/2019/04/Datatrusts-economicfunction.pdf.

Mandel, M., (2012), "Beyond goods and services: The (unmeasured) rise of the data-driven economy." *Progressive Policy Institute* 10 (2012).

Mawer, C., (2015), “Valuing data is hard,” Silicon Valley Data science blog post available at: https://www.svds.com/valuing-data-is-hard/

McKinsey, (2013), Open Data: Unlocking innovation and performance with liquid information, Report available at: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Open%20data%20Unlocking%20innovation%20and%20performance%20with%20liquid%20information/MGI_Open_data_FullReport_Oct2013.ashx

OECD, (2019), “Measuring data and data flows”, Chapter 9, Measuring the Digital Transformation: A roadmap for the future, available at: https://www.oecd.org/going-digital/mdt-roadmap-data.pdf

OECD, (2019), “Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies”, OECD Publishing, Paris, available at: https://doi.org/10.1787/276aaca8-en.

Open Data Institute, (2016), “Research: The economic value of open versus paid data”, https://theodi.org/article/research-the-economic-value-of-open-versus-paid-data/.

Open Data Science, (2019), “Data Valuation — What is Your Data Worth and How do You Value it?”, available at: https://medium.com/@ODSC/data-valuation-what-is-your-data-worth-and-how-do-you-value-it-b0a15c64e516.

Open Knowledge Foundation, (2019), “What data counts in Europe? Towards a public debate on Europe’s high value data and the PSI Directive”, available at: https://blog.okfn.org/2019/01/16/what-data-counts-in-europe-towards-a-public-debate-on-europes-high-value-data-and-the-psi-directive/.

Pinsent Masons et al., (2019), “Data trusts: legal and governance considerations”, available at: https://theodi.org/wp-content/uploads/2019/04/General-legal-report-on-data-trust.pdf.

Pope, R., (2019), “Data-sharing in government: why it’s time for a new approach”, *Medium,* available at: https://medium.com/@richardjpope/data-sharing-in-government-why-its-time-for-a-new-social-contract-7260bbe2372a.

Powles, J., and Hodson, H., (2017), “Google DeepMind and healthcare in an age of algorithms”, *Health and technology* 7.4 (2017): 351-367.

PwC, (2019), “Putting a value on data”, available at: https://www.pwc.co.uk/data-analytics/documents/putting-value-on-data.pdf, accessed on: 09.10.2019.

Raghupathi, W., and Raghupathi, V., (2014), "Big data analytics in healthcare: promise and potential." *Health information science and systems* 2.1 (2014): 3.

Raval, A., (2019), ‘BP’s Bernard Looney takes oil major into energy transition’, *Financial Times*, 6 October 2019.

Short, J., and Todd, S., (2017), “What’s your data worth?” *MIT Sloan Management Review*, Spring 2017, available at: https://sloanreview.mit.edu/article/whats-your-data-worth/.

Slotin, J., (2018), “What Do We Know About the Value of Data?”, Global Partnership for Sustainable Development Data, 2018, available at: http://www.data4sdgs.org/news/what-do-we-know-about-value-data.

Snaith, B., (2018), “Data’s value: how and why should we measure it?”, Open Data Institute blog post, available at: https://theodi.org/article/datas-value-how-and-why-should-we-measure-it/.

Spiekermann, S., et al., (2015), "The challenges of personal data markets and privacy", *Electronic markets* 25.2 (2015): 161-167.

Statistics Canada, (2019), “Measuring the value of data, databases and data science”, June 2019, PDF.

Transport Systems Catapult, (2017), The case for government involvement to incentivise data sharing in the UK intelligent mobility sector, Briefing paper.

Verhulst, S., (2019), “Sharing Private Data for Public Good”, *Project Syndicate,* online article available at: https://www.project-syndicate.org/commentary/private-data-public-policy-collaboration-by-stefaan-g-verhulst-2019-08.

Winegar, A. G., and Sunstein, C., (2019), "How Much Is Data Privacy Worth? A Preliminary Investigation." *A Preliminary Investigation (May 9, 2019)*, Harvard Law School, available at: http://www.law.harvard.edu/programs/olin_center/papers/pdf/Sunstein_1017.pdf. (Forthcoming in Journal of Consumer Policy).

About the Bennett Institute for Public Policy

The Bennett Institute for Public Policy, established in 2018, conducts high-level academic and policy research, as well as expanding the portfolio of public policy education and training offered at the University of Cambridge. The institute aims to become a world-leader in achieving successful and sustainable solutions to some of the most pressing problems of our time. **bennettinstitute.cam.ac.uk**

About the ODI

The Open Data Institute is an independent, non-profit, nonpartisan company headquartered in London. The ODI was co-founded in 2012 by the inventor of the web Sir Tim Berners-Lee and Artificial Intelligence expert Sir Nigel Shadbolt to advocate for the innovative use of data to affect positive change across the globe. The ODI works with companies and governments to build an open, trustworthy data ecosystem, where people can make better decisions using data and manage any harmful impacts. **theodi.org**

About the Nuffield Foundation

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in Education, Welfare, and Justice. It also funds student programmes that provide opportunities for young people to develop skills in quantitative and scientific methods. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics and the Ada Lovelace Institute. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation. **nuffieldfoundation.org**

The Bennett Institute for Public Policy
Department of Politics and International Studies
Alison Richard Building
7 West Road
Cambridge, CB3 9DT

Bennett Institute for Public Policy Cambridge

theODI.org

UNIVERSITY OF CAMBRIDGE

Nuffield Foundation