The valuation of data: policy implications

Draft - not for wider circulation or citation 27 December 2019 version

Contents

Diane Coyle
Stephanie Diepeveen
Julia Wdowin
Bennett Institute, University of Cambridge

Lawrence Kay
Jeni Tennison
Open Data Institute

Introduction

There is a lively debate about the value of data, but the *creation* of value from data of different kinds, its *capture* by different entities, and its *distribution*, need to be better understood. This matters for effective policy as well as business opportunities, in order to ensure that society as a whole gains from the data-fuelled changes in the economy. Yet at present there is too little distinction in the debate between different types and uses of data, and the private and public value these could create, even though the number of data transactions is growing significantly .

This report proposes an approach to understanding questions of value, and the policy implications, based on the economic characteristics of data. Its aim is to contribute to a shared understanding of the value of this newly pervasive intangible asset. By 'value', we are referring to the economic concept of 'social welfare': the broad economic well-being of all of society, including the profitability of businesses, the incomes and needs of individuals, and non-monetary benefits such as convenience or health.

This definition encompasses the value exchanges that are taking place involving public sector organisations in areas such as transport and health, with commercial deals on various terms involving patient or passenger data. We set out some key issues and principles for data policy and regulation. The ultimate aim is to ensure that there is as much creation of value as possible from data (in terms of social welfare), shared widely in society. The focus here is therefore on economic value, broadly understood from the perspective of society, not solely on commercial potential. This lens highlights some potential policy trade-offs.

Policy interest in data has two dimensions. First, government has to make policy decisions involving the value of data to the economy as a whole. These include decisions by government to invest in maintaining datasets, choices about the terms on which publicly-created data will be made available, and decisions to regulate more broadly concerning data access. The UK Treasury recently published a discussion paper pointing to the economic potential of data, but also the challenges around unlocking that potential (HMT 2018).[1] The European Commission's Joint Research Centre noted the large array of policy questions, concluding that there were no easy answers to them (Duch-Brown et al, 2017).[2] A greater understanding of the value of data would help identify where the benefits of greater investment in and sharing of data are worth the costs. Given the public good characteristics of data, it seems likely that there is considerable untapped value in enabling greater provision, access to, and joining up of data sets.

[1] HM Treasury, (2018), 'The Economic Value of Data: A discussion paper.'
[2] 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.

Second, even where it is recognised that greater sharing of data brings benefits, such as in transport, or health, those making decisions about providing data they have created need to understand the economic transaction. Although it may prove difficult or impossible to establish specific monetary valuations of certain data sets, a clearer sense of the social value is needed urgently, as many transactions involving access to public sector data are in fact already occurring. Commercial firms are eager to gain access to data held by public bodies. Yet even though these companies may develop useful and commercially successful new services, there is a risk that much potential value will be concentrated in a small number of hands, or that citizens and taxpayers will not receive a fair return from private companies using publicly provided data. These are pressing issues: following recommendations from the Hall & Pesenti AI Review,[3] the Office for AI has been investigating implementation of data trusts, which will require an understanding of how to distribute value from users to contributors. The question of market power based on data aggregation was one of the considerations for the Furman Review of competition in digital markets, which concluded opportunities for innovation and growth are being limited by a lack of access to data.[4] There is considerable debate about possible mechanisms for paying people for personal data. At the same time, there are trade-offs, particularly the need to ensure adequate incentives for investment in data, and the risks involved in storing and using data, and protecting privacy. A better understanding of data value will inform these discussions and help to shape appropriate regulation and governance, in a context of significant distrust of the uses to which individuals' data may be put by both public and private sector entities.

While a growing number of studies have investigated the value of data (reported in our separate literature review), most treat data as homogeneous. However, the value of different types of data can be very different: for example, it may be reference data, streaming data, historical data, statistical data or sensitive individual data; it may have different levels of completeness, accuracy or representativeness; it may depreciate more or less rapidly; it may be unique or commonplace; its marginal value may differ depending on context and use; and so on. This paper starts from the basic economics of data in order to develop a more nuanced understanding of how to value it using the two lenses of economic characteristics and informational content.

Many of the available empirical studies use market valuations or transactions as the basis for estimating the value of data. By definition, these valuations exclude externalities and complementarities. In effect, considering data markets as a basis for policy leaves public value on the table – for instance, the additional value that could be derived from enhanced access to enable additional uses; or the additional value from being able to combine different data sets. The (social welfare) value left untapped by failing to enable these non-market opportunities is of significant policy relevance.

---

[3] Hall, W., & Pesenti, J., (2017), 'Growing the artificial intelligence industry in the UK', *Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. Part of the Industrial Strategy UK and the Commonwealth*.

[4] 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.

This report draws on a series of expert interviews, discussions with our advisory group, feedback from conference presentations, our economic analysis, and a review of the existing literature. It touches on a wide range of challenging issues and trade-offs, and considerable work remains in order to flesh out the policy implications of the data landscape. With this caution, we conclude that it is possible to generate more society-wide economic value from data if the right set of policies can be put in place; but if the data economy is 'left to the market' there will be worse outcomes in terms of social welfare because there is a wedge between private and social incentives due to the underlying economic characteristics of data.

We set out some guidelines in the final section of the report. The key points concern the need to consider in detail, in different contexts, the access rights different organisations or individuals have to certain data, and establishing a trustworthy institutional framework for managing, monitoring and enforcing the terms of access. Asymmetries of information mean that contracts for data use are incomplete, and the regulatory framework should recognise this, particularly that schemes for sharing data in a regulated way change the returns on investment in collecting and cleaning data, and investing in complementary skills and assets. There are also some unavoidable trade-offs that will require policy choices.

The distinctive economic characteristics of data

There are several existing taxonomies of data aiming to delineate characteristics relevant to valuation. Some are presented in the table below.

By characteristics	**By origin**	**By usage**	**By feature**
OECD 2013[5]	OECD 2013[6]	Sweden National Board of Trade 2014[7]	Nguyen & Paczos 2018[8]
•Sensitivity •Subject •Purposes •Context •Identifiability •In/directly collected	•Provided •Observed •Derived •Inferred	•Corporate •B2C •Human resources •B2B •Technical	•Public or private •Proprietary or open/public domain •Personal or not •User created/machine generated/administrative •Actively or passively created

[5] OECD (2013), "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value", *OECD Digital Economy Papers*, No. 220, OECD Publishing, Paris, https://doi.org/10.1787/5k486qtxldmq-en.
[6] Ibid.
[7] The National Board of Trade (2014) "No Transfer, No Trade: The Importance of Cross-Border Data Transfers for Companies Based in Sweden," The National Board of Trade, p8
[8] Nguyen, D. and Paczos, M. (2019), "Measuring the economic value of data and data flows", presentation at OECD Working Party on Measurement and Analysis of the Digital Economy, Paris, 7 May.

However, data is an intangible asset with distinctive economic characteristics, which do not map onto these taxonomies. This economic lens is the first one we apply to consider the value of data.

Most importantly, data is *non-rival*: unlike many conventional goods (such as apples) or assets (such as machine tools), many people can use the same data at the same time without it being used up. It is, technically speaking, either a *public good*, or a *club good* when access to it is excluded by technical and/or legal means. Data is therefore shared to varying degrees, or its use is licensed. It is not best thought of as owned or exchanged. Our interviewees were unanimous in agreeing that 'ownership' is an inappropriate concept for data (and that characterising data as 'the new oil' is similarly misleading).

Data often involves *externalities.* These are often positive, such as additional data improving predictive accuracy, or enhancing the information content of other data.[9] In these cases data often gains its value from being combined with other data. For example, one person's health data gains much of its value from comparison with aggregate statistics, such as the distribution of cholesterol levels in the population or average blood pressure, and other data based on research about how these are linked to health outcomes. There may also be negative externalities, notably the compromising of individual privacy.[10]

Although there is a strict legal definition of personal data in the EU under GDPR, which means it has to be treated differently from other types of data (implying different costs and risks), our interviewees by and large considered it would be impossible to define with any precision an economically meaningful category of 'personal data'. For although individuals provide considerable amounts of data about themselves, which may be sensitive or private, the valuable information content often lies in aggregation or in comparison with data provided by others. The information is co-produced by individuals and by companies, with individuals creating (positive or negative) informational externalities for each other.[11] Consequently, focusing questions of value on the data provided by individuals overlooks the allocation of value created thanks to the externalities and complementarities.

Along with this public good character of data (in the technical sense of non-rivalry), the externalities mean that market mechanisms are unlikely to deliver socially optimal outcomes, producing too little and/or charging too much where there are positive externalities, and vice versa. We are probably in a situation where both are true: there is over-production and commercial use of some types of data raising privacy concerns; and also under-production and use in contexts where the commercial opportunities may not be so obvious (or may be limited to the aggregation of data across consumers or across activities occurring within individual firms) but the potential public value is large. Together,

[9] Charles Jones & Christopher Tonetti, 'Nonrivalry and the Economics of Data', Stanford Business School Working Paper, August 2019.
https://www.gsb.stanford.edu/faculty-research/working-papers/nonrivalry-economics-data
[10] Acemoglu et al, 'Too Much Data', MIT Working Paper, September 2019
https://economics.mit.edu/files/17760
[11] See also Dirk Bergemann & Alessandro Bonatti (2019), 'The Economics of Social Data: An Introduction', Cowles Foundation Discussion Paper 2171R, September.

the non-rivalry and externalities mean there is a wedge between the private value of data and public value (social welfare in economists' language). When there are positive externalities, and information content comes from aggregation, too little data will be provided for use, from the perspective of society, as it can be difficult for whoever incurs the cost to capture the benefits of it. In either case, market transactions alone will not lead to the best social outcomes; a strategy of public investment and regulation is essential.

The distribution of value is also affected. The identity of the beneficiaries of insights from data influences its total potential value, how that value is likely to be distributed, and the likelihood of investment in that data. For example, data about purchasing decisions may be valuable to advertising agencies; it could either boost their profits or reduce costs for advertisers. Data about disease contagion may be valuable to public health professionals; it may increase the impact of money spent on public health measures and reduce sickness and death. Data about purchasing decisions might attract a high market price because of the direct economic benefit to advertising agencies, while that about disease does not.

Aggregated value may often be greater than the sum of individual values, but sometimes there can be *increasing returns* to gathering more data, and sometimes *diminishing returns*. This is determined by context and use. Sometimes data is needed to build a predictive statistical model so at some point, after a period of increasing returns, diminishing returns to additional data points set in. Nevertheless, data holders with market power may continue to accumulate data as a means of cementing their economic rents and using data as a barrier to entry. In other cases, such as certain mapping and traffic applications, granular real time data are useful and more data points will continue to add value. Data can also *depreciate*, losing value, at different rates depending on context and use - slowly for population health data, much faster for data concerning real time traffic flows supporting in-car navigation systems. So the volume of data, its granularity, and its speed - as well as its accuracy - will all have implications for value but in varying ways depending on context.

As well as the *context*, discussed further below, the *consequences* of data use affect value. Additional data will be more valuable in highly consequential situations. Contrast the potential life and death consequences of autonomous vehicles with the consequences of badly targeted advertising. The *use values* differ widely.

Often individual sources of data will have considerable *option value*, or in other words might become valuable if new questions, not yet thought of, can be answered in future. The consensus among our interviewees was that many of the organisations accumulating data have been doing so because of potential rather than actual uses. The EU's GDPR legislation rules out the accumulation of individuals' data for other than specified reasons but this may rule out potential for innovation; significant innovations usually derive from new questions rather than new answers to old questions. While the legislation does not formally require individuals' permission to be sought for any new use of their data, many companies are currently taking an ultra-cautious approach. It may take some time for GDPR to be fully understood and tested.

Data also involves costs and risks; it can be a liability as well as an asset.

Investment in the collection and cleaning of data often has a *high up-front cost and low marginal cost* (like other digital or intangible goods and assets). Up-front costs might involve investment in hardware (such as sensors), software, data modelling or standardisation, and in developing processes for collecting and maintaining data. The return on the investment will depend on the use of the data. This use value is likely to exceed the marginal cost, particularly when the collection of data is highly automated, for example generated through the delivery of a digital service or from sensors. Secure storage of data - necessary when data is sensitive - involves costs, and while storage has become relatively cheap, the risk of data breaches has increased. There are additionally reputational and financial risks (including fines) associated with security breaches or data misuse. When data collection is laborious, for example involving surveying of people, organisations or the physical environment to create maps or registers, the costs of maintaining data can also be high.

Finally, capturing the value from data will often need specific *capabilities* (e.g. data science and analytical skills, management know-how) or *complementary investments* (e.g. software, other capital equipment). Our interviews consistently indicated that a lack of capabilities is a major barrier to capturing the potential value from data use.

Together, these considerations imply the following issues and potential valuation methods:

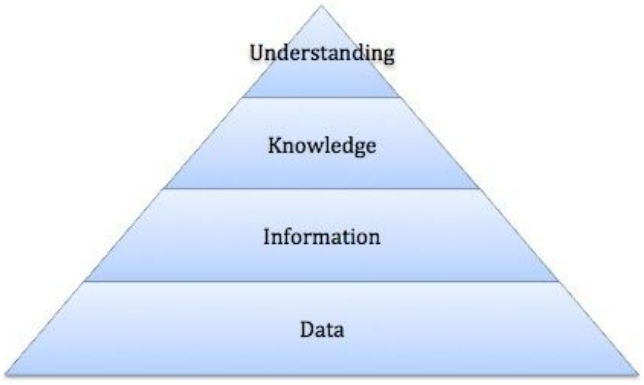
Characteristic	Issues	Evaluation
Diminishing/increasing marginal returns?	How granular is the necessary data? How much data is needed for prediction models? Is the holder using data accumulation as a source of market power?	Accuracy of predictive models Innovations and quality improvements in services Monopoly rents - profitability, absence of new entry
Externalities	Does additional data, or aggregation, sharing/open data, or joining different data sources add information? Does it compromise privacy?	Innovations and quality improvements in services Contingent valuation methods
Optionality	Does gathering more information provide scope for future process or quality improvements or innovation?	Real options methods
Consequences	Are decisions made using the data highly	Value at Risk methods

	consequential?	
Costs	What costs need to be covered - data acquisition, data cleaning, storage, skills and capabilities, governance? What are the contingent costs - security breaches, loss of sensitive information, reputational damage, fines?	 Harm to identified individuals (eg if later defrauded), loss of commercial confidentiality Risk assessments

Subject, context and use

The second lens to apply to data value concerns its *information content* - illustrated by the classic information pyramid, Figure 1 below: information enables people, firms and government officials to make better decisions, depending on their objectives. Context matters because the value of data is not related in any simple way to its volume (records, bits etc).

Figure 1: The Information Pyramid



The subject or information content of data determines how useful it will be. A number of characteristics shape this.

Some of these reflect use value. Data can be about people (such as demographics, behaviours, and relationships), about organisations (such as their types, activities and business relationships), about the natural environment, built environment or manufactured

objects. It can be used to make decisions that affect us economically - such as about purchases or investments; our environment - such as our energy and transport use; or our lives - such as our health, education or engagement with society.

The *generality* of data determines how many decisions the data is useful for. Some data might be only valuable for a few purposes and other data useful in a range of different scenarios. For example, labelled retinal scans might only be useful for creating diagnostic systems for eye diseases whereas geospatial data might be used for things as varied as navigation, understanding the density of services offered to different communities, or predicting the impact of floods.

The *temporal characteristics* of data also determine how it can be used. Data can be:

- Forecasts that predict what will happen in the future
- From the present or recent past
- Part of the historical record
- A backcast that estimates what happened in the past

Data can be used by people and organisations taking different kinds of actions:

- Planners - acting to affect our prepare for the short/medium/long term future eg city planners, children choosing schools/subjects
- Operators - acting to deal with the present, eg doctors in A&E, commuters deciding what route to take home
- Historians - acting to respond to something in the past eg police investigating a crime, tax collectors

The utility of different temporal characteristics for these different people are shown in the following table:

	Planners	Operators	Historians
Forecasts	Most valuable	Near future potentially valuable if it helps prioritise	Not valuable
Current/recent past	Valuable to feed into prediction engines	Most valuable	Valuable only in so far as it provides an anchor for understanding what happened in the past
Historical record	Valuable to create and validate prediction engines	Only valuable in so far as informs current decision making	Most valuable
Backcasts	Valuable when historical record is	Not valuable	Potentially valuable if it supplements or

	missing, to supplement existing data in the generation of prediction engines		supports the historical record

Quality

Quality is frequently cited as an important characteristic of data. However, it is important to note that the quality needed depends on what it is used for. Higher quality data reduces uncertainty or reduces the risk that decisions based on it are incorrect; knowledge of known issues with quality can help with assessing that risk. Data quality is typically described in terms of characteristics such as completeness, accuracy, and timeliness:

- *Completeness* is an assessment of what proportion of reality a dataset represents. This can include its spatial and temporal extent as well as being influenced by sampling and biases in data collection.
- *Accuracy* is an assessment of the correctness of the information made available in the dataset. Accuracy can be influenced by the method of data collection, with more automated mechanisms being more accurate.
- *Timeliness* is an assessment of the delay between the time period the data is about (its temporal extent) and when it is available. Timeliness is particularly important for data being used in an operational context, and for data that relates to the recent past and forecasts of the near future.

The lower the completeness, accuracy and timeliness of data, then in general the greater the marginal returns on additional data being incorporated into the dataset.

Sensitivity and personal data

Data can be sensitive for a number of reasons, including revealing information about individuals or organisations, or about physical assets that might be susceptible to attacks or disruption. Sensitive data will normally have restrictions on access to protect the people, organisations or things it is about. However, sensitive data is usually thought to be vital for personalising or customising a product or service. The requirement to protect sensitive data means that collecting and storing it will have additional costs.

Interoperability and linkability

As discussed above, the value of data frequently arises from it being brought together with other data. There are two characteristics that relate to the ease of combining datasets:

- *Interoperability* relates to the use of data standards when representing data, which means that data relating to the same things can be easily brought together.
- *Linkability* relates to the use of standard identifiers within the dataset that enables a record in one dataset to be connected to additional data in another dataset.

Visibility and excludability

Once collected and accessible, data is non-rival. However, some types of data can be naturally excludable while others are not. Categories for data based on its excludability include:

- *Environmental data* is data collected about the environment. Anyone can see data that arises from the environment, so anyone (with the right sensors) can collect it. It is hard to exclude environmental data from those who can afford to collect it. Examples include: geospatial data, rainfall, satellite data, air pollution, roadworks, and data that is public on the internet such as tweets or LinkedIn profiles.
- *Administrative data* is data that is collected as people interact with public or private services. Unless it is explicitly shared, this kind of data is only naturally visible to those providing the service eg tax returns to tax offices, shopping carts to retailers, patient records to healthcare providers.
- *Planned data* is data about planned activities. This information is invisible except to those doing the planning and is therefore very easy to exclude. Examples include budgets, roadwork schedules, or bus timetables.
- *Predicted data* is data about what might happen in the future. This may be data anyone can create, if they have access to enough data to create reasonable predictions and capability to create predictive models. Examples include voting outcomes, weather forecasts, or stockmarket predictions.
- *Historic data* is data about historic events. This data is only accessible to those who were there or who recorded it, although it may be reconstituted through backcasting in a similar way to predicted data. Examples include my browser history, historic members of parliament, actual bus times.

Accessibility

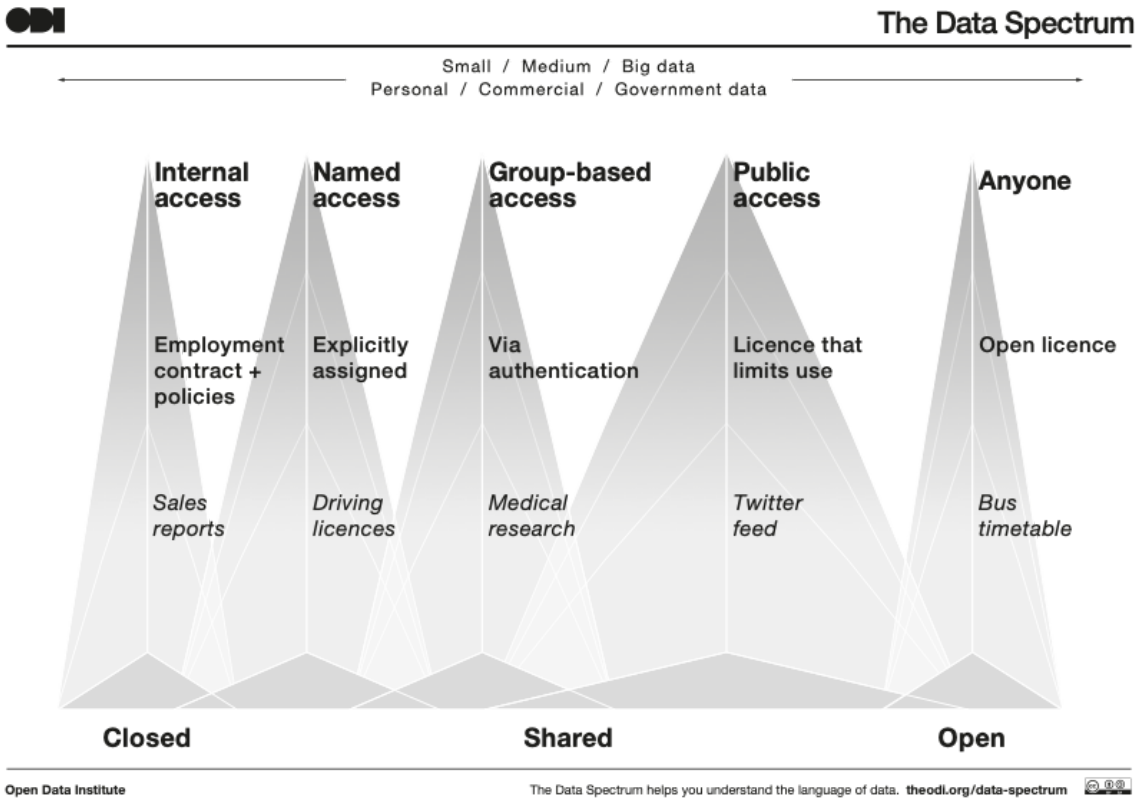
The useful distinctions above do not, however, capture the role of access to data in unlocking its wider value to economies and societies. To move toward an alternative approach, we begin by describing in more detail the Data Spectrum. Data, although inherently non-rival, can be closed, shared, or open:

> If access to data is restricted, its uses are limited to what that organisation can do with it.
>
> If it is shared with a select group of people - a *club good* - its uses and analysis can be wider, perhaps creating more value.
>
> If data is shared openly - a *public good* - anyone can use it.

The creation of wider social value through greater openness is limited by the sensitivity of the data - sharing more widely can cause harm to the people, organisations or environment the data describes - and by incentives to invest, as firms might be reluctant to spend money on collecting and controlling the data if others are going to capture the benefits.

The Data Spectrum (Figure 2 below) shows some of the access conditions determining whether data is a private, shared, or public good. Access conditions can be determined by technology, licensing or terms and conditions, and regulation. For example, an authentication process for parties wanting to access data on a medical research project restricts availability, which safeguards sensitive data and perhaps raises value creation potential by enhancing incentives for investment in long-term research. On the other hand, a lack of interoperable standards or restrictive licensing will reduce potential investment and value creation.

Figure 2: The Data Spectrum



Access conditions can be determined by how many overlapping use rights might apply to the data. If data is collected about someone who has a high degree of control over it, they have something akin to a unilateral property right and may be able to demand payment for others to use it. Where data is collected about many people at once, and they have rights to control sharing or use of that data, there will need to be a process of negotiating claims to control before it can be used.

Data about individuals can be found at both ends of the Data Spectrum, and rights and access affect its value.[12] Information about the consumption, movement, and work habits of a person can be valuable to advertisers, particularly when similar data is collected about lots of other people. But the risk of harm from sensitive insights about that behaviour being released inappropriately has long motivated data protection laws around the world and hence barriers to access.[13]

The current UK legal framework

The trading of data, and the distribution of value arising from it, are founded on legal rights. This section describes the basics of the current rights framework in the UK.

Intellectual property rights and licensing

When an organisation or individual creates an intangible asset, such as documents, code or data, they automatically have intellectual property rights in that asset. The most important category when it comes to data are *copyright* - rights over assets generated through creativity - and *database rights* - rights over datasets arising when significant effort is invested in creating or maintaining that data. There are no intellectual property rights in plain facts; these only arise when facts are arranged into databases. Unlike copyright, *sui generis* database rights are only defined in a few countries, mostly in the European Union, and they exist in the UK by virtue of The Copyright and Rights in Databases Regulations 1997.[14] These have a term of 15 years, but as this is extended when substantial alterations are made to the database, there is essentially no termination date for database rights in data constantly kept up to date.

These intellectual property rights limit what other people can do without explicit permission from the rights holder. Permission to use data and other content can be explicitly provided either by licensing it or by dedicating it to the public domain, which means waiving IP rights in the asset.

Licences to use data can be generated on a case-by-case basis, through negotiation between the rights holder and the licensee. However, more typically a rights holder will have a fixed licence that it applies to particular data. There are also standard licences, most

[12] Open Data Institute (2019) Anonymisation and open data: An introduction to managing the risk of re-identification, https://docs.google.com/document/d/1CoXniaTnQL_4ZyQuji9_MA_YCEElQjx4z1SEdB08c2M

[13] Although the EU's Digital Single Market strategy has distinguished non-personal data – such as on transport timetables – from personal data in an attempt to simplify transactions. European Commission (2019) 'Free flow of non-personal data.' https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data; Eur-Lex (2018) 'Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union' https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1546942605408&uri=CELEX:32018R1807

[14] http://www.legislation.gov.uk/uksi/1997/3032/made

notably open licences, which are adopted by multiple licensors. Standard licences reduce the transaction and legal costs involved in setting up an arrangement to use data.

This legal framework thus provides for the holders of database rights to charge for:

- The IP rights themselves, i.e. to transfer both a database and the rights to use that database and determine who else uses it, to someone else;
- A one-off licence to use a particular version of a dataset;
- A continuous or recurring licence, where the licensee will pay a subscription to retain access to up-to-date data.

Licences can and frequently do contain clauses that limit what licensees can do with data, in particular to protect the licensor from losing revenue if licensees make the data available to third parties. Licences may limit the ability of the licensee to sub-license, allow this only in return for additional royalties, and may claim additional rights over data derived from the originally licensed data.

Intellectual property rights in public sector information

The intellectual property rights in public sector information (PSI) - that is information generated by a public body in the course of delivering on its public task - is held by the Crown. Those rights are administered by the Queen's Printer, within The National Archives, through the Government Licensing Framework.[15] Under this framework, most PSI that does not contain personal data is licensed with the Open Government Licence (OGL). Public bodies can only license data differently if they are given a delegation of authority allowing them to do so.

The Reuse of Public Sector Information Regulations 2015 constrains the ways in which PSI can be licensed, in particular ensuring that no exclusive licences are granted (which would prevent the public body from granting a licence to other reusers) and ensuring no one is given preferential terms. Access to some data created by public bodies can also be requested through the Freedom of Information Act 2000 and the Freedom of Information (Scotland) Act 2002. The Freedom of Information Act 2000 was amended by the Protection of Freedoms Act 2012 to ensure that public bodies provide clarity about the licensing of any data requested through the act. The Environmental Information Regulations 2004 also provides a mechanism for accessing environmental information, including from some private sector bodies such as water companies who are delivering a public service.

These rights and responsibilities are regulated and enforced by the Information Commissioner's Office (ICO).

[15] Government Licensing Framework. http://www.nationalarchives.gov.uk/information-management/re-using-public-sector-information/uk-government-licensing-framework/

Data protection rights

While they are alive, people have a set of rights over data that is collected about them. These rights are enshrined in the Data Protection Act 2018, which maps the General Data Protection Regulations (GDPR) into UK law.

Market-based methods of data valuation

Although the economic characteristics of data mean it is unlikely that market-based transactions give a complete picture of the value or potential value of data, a growing number of studies and approaches use market prices to estimate value. These can be divided into broad categories: stockmarket valuations, and income-based or cost-based approaches.

One approach is to compare the stockmarket valuations of companies that are and are not data-intensive. For example, a report by PWC finds that stockmarket 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.[16] There are some striking examples of how effective data use can affect corporate performance. For instance, BP has a 10-year $1.2bn contract with Palantir to integrate data across its businesses. One early result is a digital model of BP's entire production system which can optimise the oil's most efficient route, using data to speed the flow and increase production by 30,000 barrels a day.[17]

An alternative approach, taken in number of recent papers and reports, is income-based, using "an estimate of future cash flows to be derived from the asset."[18] The data value chain (Figure 3) visualises this approach.

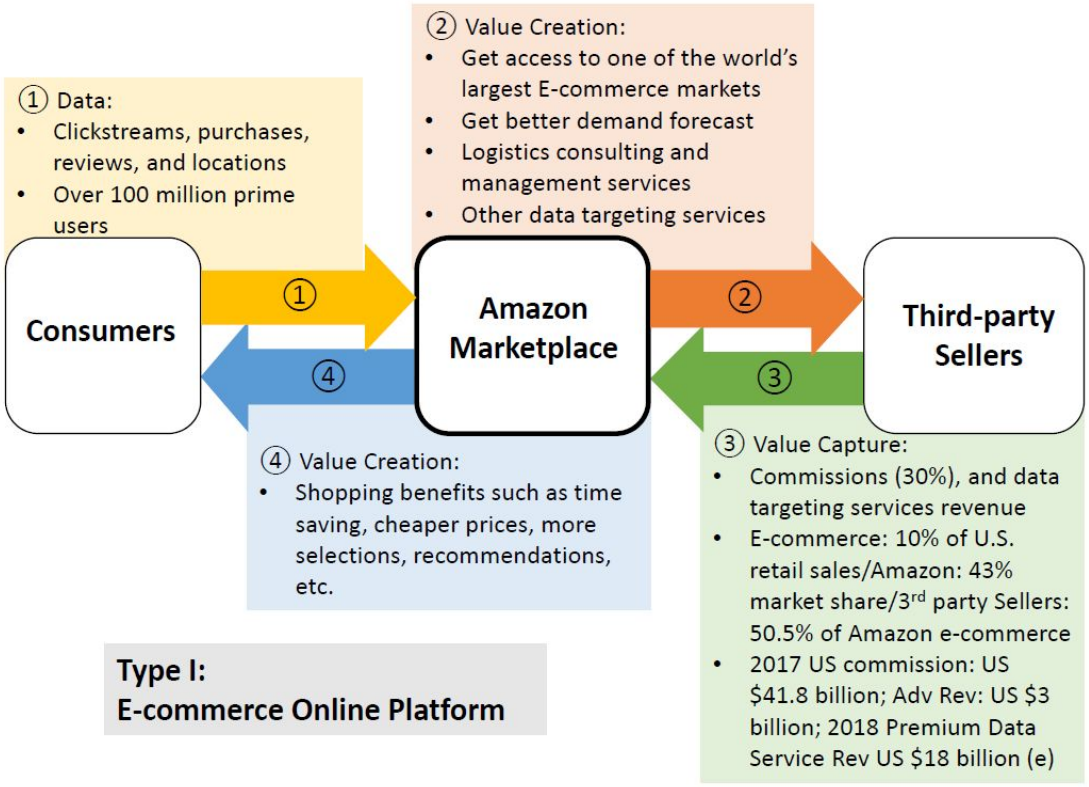
Figure 3: Data value chain

Mawer 2015

[16] 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.
[17] Anjli Raval, 'BP's Bernard Looney takes oil major into energy transition', Financial Times, 6 October 2019.
[18] Mawer, C., (2015), "Valuing data is hard," Silicon Valley Data science blog post available at: https://www.svds.com/valuing-data-is-hard/. See also Carol Corrado (2019), 'Data as an Asset', presentation at EMAEE 2019 Conference on the Economics, Governance and Management of AI, Robots and Digital Transformation; and Maria Savona (2019), 'The Value of Data: Towards a Framework to Redistribute It', SPRU Working Paper 2019-21 October.

Li et al (2019) consider data value in the context of value chains for several different business models, such as e-commerce marketplaces, search, or matching platforms. Many of these involve a direct monetary benefit to the company accumulating data, an indirect monetary benefit to suppliers and advertisers who subsequently make more sales, and a non-monetary benefit to final users who get free online services (as well as perhaps more choice or lower prices). However, the authors observe that the income-based approach is limited because a data-driven business model, embedded in the organisation's capabilities, can create additional value beyond that generated by the chain of transactions. Thus the big data-driven platforms effectively capture much of the social value of the data they have accumulated. For example, in Amazon's case the platform is able to take advantage of the feedback loops its business model creates (Figure 4).

Figure 4: Value creation in e-commerce platform



Li et al (2019)

Research into users' contingent valuation (willingness to pay/willingness to accept) suggests they place a high value on the free online services they can access in return for their provision of data.[19] However, the character of the exchange (and market power/profitability of the large data-accumulating companies) has led to some debate about whether or not

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

individuals should be paid for data they provide (Arrieta Ibarra et al., 2017).[20] This model has emerged in the case of some health data start-ups, which act as platforms matching data from patients who sign up with interested pharma companies.[21] Any such payments to individuals would be small, however, compared to the externalities created by aggregating data - which is another way of restating the limitations of market-based (monetary) transactions as a basis for valuing data.

The assessment of future income or profits can change substantially as new opportunities emerge. Technological innovations enable new ways of using data that can revalue it. For instance, 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.[22] The value of any asset depends on an estimate of future returns, so this characteristic is not unique to data assets; however, the barriers to new uses of data may be lower than in the case of other assets. Market prices at any moment in time are unlikely to include the full option value of the data.

Another limitation to market-based methods of valuation is that there are not many 'thick' data markets with a sufficient number of buyers and sellers to ensure that the transaction prices are closely related to fundamental economic value. Monetary transactions do take place, with an active landscape of data broking companies selling data about individuals for marketing purposes, and indeed a market in illegal transactions for stolen data. There are thousands of data brokers offering for sale different types of data on individuals or companies. However, as these data markets are complicated, non-transparent, and increasingly concentrated, the prices of transactions in them do not seem to be a sound basis for valuation.[23] A market study by the UK Competition and Markets Authority expressed concern about whether consumers are getting a fair deal in the data-driven online advertising market.[24] Data brokers do not post prices, and there is a wide range for estimates of the value of personal data to businesses involved in advertising-based models or digital marketing. Estimates based on prices posted on the dark web, where hacked data is sold, range from around £1 to over £200.[25]

The alternative cost-based approach is used in estimating the aggregate value of data to the economy in the national accounts, as there are relatively few market sales of datasets, with most being generated within the business in the process of providing other goods and services. The figures currently used in the national accounts are defined to reflect the costs to businesses - mainly labour costs - of preparing data in a useful format, but not of

[20] Arrieta Ibarra, I., Goff, L., Jiménez Hernández, D., Lanier, J., & Weyl, E. G., (2017), 'Should We Treat Data as Labor? Moving Beyond 'Free'.' *Moving Beyond 'Free' (December 27, 2017). American Economic Association Papers & Proceedings*, *1*(1).
[21] 'Patients take control of their medical data', Sarah Neville, *Financial Times* 23 April 2019.
[22] Arrieta Ibarra, I., Goff, L., Jiménez Hernández, D., Lanier, J., & Weyl, E. G., (2017), 'Should We Treat Data as Labor? Moving Beyond 'Free'.' *Moving Beyond 'Free' (December 27, 2017). American Economic Association Papers & Proceedings*, *1*(1).
[23] Federal Trade Commission, 'Data Brokers - A Call for Transparency and Accountability', (2014).
[24] https://www.gov.uk/cma-cases/online-platforms-and-digital-advertising-market-study#interim-report
[25] https://www.moneywise.co.uk/news/2018-03-21%E2%80%8C%E2%80%8C/how-much-your-data-worth-hackers

purchasing or producing the underlying data in the first place. There is a debate among statisticians now about whether this approach is too limited, given the explosion of data gathering and use. One approach would be to treat firms' creation of digitized data as investment in an asset, which would increase GDP compared with the current treatment, and in turn would continue to be measured in terms of the cost of creating the data. An alternative is to see data as generated by households and provided as a barter with businesses in return for free services; this approach would specifically apply to advertising-funded social media and search companies so its dependence on a particular business model is a drawback. Organisations can also use a cost-based approach to valuing the data that they collect, either by looking at the investment they put into collecting it, or by assessing the cost of replacing that data with something equivalent. The latter approach implies that unique data is priceless, whereas the value of data from things that can be observed in the environment (eg satellite imagery) is diminishing as the number of alternative sources increases.

Some of the research on market valuations illustrates the distinction between private and public value. For example, one study found that monitoring drivers prompted them to drive more safely, while the monitoring data enabled their insurance company to make a higher profit while reducing the premium charged to safer drivers. However, requiring the insurance company to share the data would have reduced its incentive to invest in the monitoring and data collection scheme.[26]

This highlights the broader point that the regulatory environment changes market valuations. The market transactions currently observed are not capturing a fundamental reality; rather, the market value of data is endogenous, depending on policy choices. There are likely to be many trade-offs between creating private incentives to invest in collecting and using data, and capturing social benefits. The value of data will depend on the societal trade-offs, analogous to those in the domain of intellectual property where there is a trade-off between the private incentive to invest in innovation created by the temporary monopoly provided by patent or copyright and the social benefit of ensuring wide access to innovations as quickly as possible. The heated debate about the economics of intellectual property in the digital economy suggests that the policy choices will be no easier in the case of data.

Existing non-market estimates

Market valuations thus provide useful information but do not capture the full social value of data. A number of studies (in addition to those cited above concerning the value of 'free' services) have provided estimates of data value going beyond market transaction values, mostly using contingent valuation methods.

A 2013 study of the impact of opening up Landsat data using this methodology estimated a value of £2bn/year, based on surveying different groups of users to estimate the average

[26] Jin, Y., and Vasserman, S., (2019), 'Buying Data from Consumers', NBER Working Paper 2019.

monetary benefit to each group.[27] This did not include the additional value to additional users provided with services based on Landsat data.

In a 2017 study of TfL's open data approach, Deloitte evaluated the cost savings and incremental value to three groups – passengers and other road users; the London economy as measured by job creation and commercial use of the data by firms; and TfL itself – generated from TfL's £1m a year investment in publishing open data.[28] For passengers, they estimated £70m-£90m/year cost savings through less time being wasted in adjusting routes in light of new information; they also highlighted value arising from increased use of the public transport network particularly by those with accessibility needs (£5.1m/year) and healthier lifestyles due to increased cycling and walking. For the London economy, value arose from new companies using open data, amounting to £14m a year in GVA and the generation of 500 direct and 230 indirect additional jobs. For TfL, value was reflected in £1m costs saved from customer support services they would otherwise need to provide directly.

A 2019 report on the value of Companies House Data included a valuation for intermediaries such as credit reference agencies who use Companies House data as an input to their own data products and services as well as people and organisations who access the information directly from Companies House.[29] Intermediaries attributed £23m/year of their revenues and £5m/year of their costs to their use of Companies House data. They did not attempt to quantify the impact of this data being absent, but described costs associated with removing functionality from their products and services or collecting relevant data from businesses directly themselves.

When 2,416 individuals were asked how much they valued their data privacy, Angela Winegar and Cass Sunstein found that willingness to pay for privacy was low (an average $5 a month) but willingness to accept loss of privacy was a more substantial $80 a month.[30] Both figures are higher than the amount indicated by calculations such as Facebook's average profit per active user (about $2) or - a different type of benchmark - the amount per individual implied by fines for data breaches ($125 per person implied by Equifax's 2019 fine from the US FTC; £0.005 per affected user implied by the UK ICO's 2019 £500,000 fine on Facebook).

While the different approaches each have limitations, all these studies highlight that valuation of data in a wider societal context needs to consider different groups:

---

[27] Miller, H.M., Richardson, Leslie, Koontz, S.R., Loomis, John, and Koontz, Lynne, 2013, Users, uses, and value of Landsat satellite imagery—Results from the 2012 survey of users: U.S. Geological Survey Open-File Report 2013–1269, 51 p., http://dx.doi.org/10.3133/ofr20131269

[28] Deloitte, 2017, "Assessing the value of TfL's open data and digital partnerships" http://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf

[29] UK Government (BEIS) "Companies House data: valuing the user benefits" https://www.gov.uk/government/publications/companies-house-data-valuing-the-user-benefits

[30] 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). Advocates of data privacy rights challenge the very notion of a price for privacy, reflecting the broader debate about the validity of contingent valuation methods that seek to put monetary values on intrinsic goods.

1. The costs and benefits (or risks and options) to data stewards of collecting, using and sharing data;
2. The costs and benefits to intermediaries with whom data is shared, and the wider economic impact of the activity of those organisations (for example in providing jobs). Each of the above studies highlights that products, services and entire businesses can be brought into being due to data being available to them. For intermediaries, attributes of data, such as quality and interoperability, are important for reducing costs and risks, as are aspects of their relationship with the data steward, such as receiving notifications of changes;
3. The costs and benefits to end users or consumers who use the products and services provided by intermediaries.

Creating value through open and shared data

A dataset holds information which needs to be analysed before it can be used in a product or service to meet demand in a given context. Making data closed, shared, or open means changing the range of people who might analyse the digitally stored information, be able to turn it into a product, or use it in different contexts.

The public good character of data and the prevalence of positive externalities create a presumption that more open access to certain types of data will increase social welfare. However, there are several trade-offs to consider.

First, there is a trade-off due to the need to incentivise investment and innovation and to cover ongoing costs of maintaining data securely. This is similar to the well-known trade-off in intellectual property, where patent or copyright protection restricting access is needed to create an incentive for investment in discovery and innovation to occur in the first place, but at the same time limits the potential social welfare benefits of a new service or product.

This trade-off is most direct for organisations whose purpose and business model centres on data collection and maintenance as opposed to those generating exhaust data as a consequence of their activities. In the former case, the cost of generating data has to be met. In the latter case decision-making about the investment is driven by the benefits of providing the data-generating service.

However, the need to provide more access to data can also disincentivise investment in products and services that use data. Exclusive access to data can enable firms to gain a market advantage for the services they offer, such as providing a more personalised service to their customers. If that data is also available to other companies to provide an improved service, that private advantage is lost although social benefits will likely be enhanced.

For most organisations, providing access to certain types of data (individual or sensitive) is an additional cost which may be difficult to meet. They are therefore likely to underinvest in

the provision of such data, if they do not charge for it. This may be particularly the case for public sector organisations with limited budgets for this purpose or alternatively if they are required to recoup some of their costs from charges. .

A second trade-off applies to data that is personally or commercially sensitive. Individuals will want to limit access to certain types of identifiable information about themselves. Companies will not want to share data that will help their rivals. So increasing access to such data has attractions but also involves complexities, including avoiding negative privacy-intruding externalities.

A final trade-off concerns the requirement for an evidence base before investing in data or providing access to data. The option value of data - the fact that it is hard to predict how data might be used by other organisations or in the future as technologies change and other data becomes available - means there is inherent uncertainty when making cost/benefit trade-offs around investments in improving the quality of or access to data. Even the work needed to understand the potential realisable value of data in order to reduce this uncertainty is complex and costly in and of itself. This can lead to under-investment as organisations await greater certainty.

Moves toward open *government* data have been motivated by the publishing of public sector information as a public good. That motivation comes from expectations of the value that it creates, in terms of transparency and accountability of democratic institutions, and the stimulation of innovation and economic growth.[31] Allowing citizens to have more access to the information that the government holds can give them the opportunity to make more informed decisions, while also analysing the data with skills that might not be available, or useful, to the government. The government also collects substantial administrative data whose wider use could also enable better decisions and improved services. As with any open data, government data can be combined with private and shared data.

Estimates for the value of open data as a percentage of GDP have ranged from 0.08 percent to 7.19 percent, derived from different mixtures of sectors, countries, types of data, potential benefits, and other factors (Figure 5).[32] A recent OECD report cites a range of 1 per cent to 2.5 percent of GDP.[33] The range of estimates may be partly caused by the lack of research

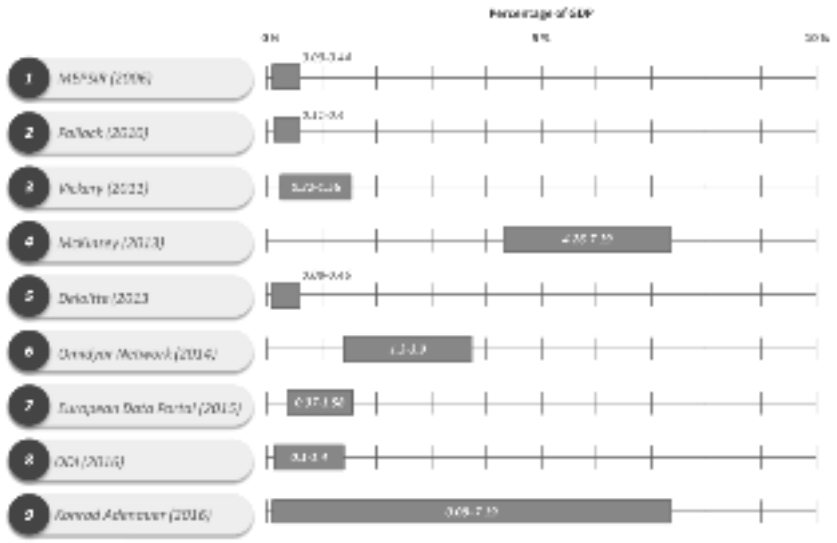
[31] '[Open government data]...is, data which traditionally originates from governments, is created or used during the business of governing, or is created or published at the request of governments', see Davies T Smart metering equipment could potentially be used to collect property information, such as temperature or humidity measurements, to spot where there are health risks to vulnerable people.etal (eds) (2019) The State of Open Data, p7. 'Public sector information (PSI) is information produced by central and local government or any other public body', see The National Archives 'About PSI.'

[32] European Data Portal (2017) 'Analytical Report 9: The Economic Benefits of Open Data', https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf, p17

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

into the effects of open data in comparison with properly delineated counterfactuals, as noted by the 2018 Open Data Barometer report.[34]

Figure 5: the value of Open Data as measured by different studies[35]



The debate about the value of open public sector data has led to consideration of which datasets might be of most value.[36] The European Commission sees the value of public sector information as determined by its potential to create economic and other benefits; the potential for the creation of innovative services; how many people can use it; the scope for revenue; the possibilities for re-combination; and the effects on public undertakings.[37] These suggest six types of government data that have the most value: geospatial, earth observation and environment, meteorological, statistics, companies, and transport.[38]

[34] Open Data Barometer (2018) Open Data Barometer: Leaders Edition, from Promise to Progress https://opendatabarometer.org/?_year=2017&indicator=ODB

[35] European Data Portal (2017) 'Analytical Report 9: The Economic Benefits of Open Data', https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf p18; also see Kuzev P (ed) (2016) Open Data The Benefits, https://www.kas.de/einzeltitel/-/content/open-data.-the-benefits1

[36] Open Knowledge Foundation (2019) 'What data counts in Europe? Towards a public debate on Europe's high value data and the PSI Directive' 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/

[37] Open Knowledge Foundation (2019) 'What data counts in Europe? Towards a public debate on Europe's high value data and the PSI Directive' 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/

[38] Open Knowledge Foundation (2019) 'What data counts in Europe? Towards a public debate on Europe's high value data and the PSI Directive' 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/

However, creating public sector information and making it openly available is costly. Although the amount spent on official statistics and other public data is low in per capita terms - and, as noted above, there is almost certainly too little provision - governments often consider requiring payments for publicly-held data to help cover the costs. What's more, EU legislation has often been interpreted to require that all users are charged the same amount, from individuals to big corporations, even though the economic calculation in terms of marginal cost and benefit faced by different types of user differ greatly. Yet requiring payments for public sector data can impede its use and hence the value that can be derived from it; research for the Open Data Institute found that making the most useful public datasets open would create 0.5 percent more GDP growth per year for the British economy than making users pay for access to the data.[39]

Private sector organisations can also open the data that they hold.[40] In the development of artificial intelligence, firms can adopt business models that make their algorithms and the data they control more or less open, affecting how easy it is for them to collaborate with others and limiting the costs they face in managing large datasets.[41] However, many large datasets are held by the private sector and are far less open than public data. At present there are relatively few incentives or legal requirements for private sector companies to share their data (although in the UK the Digital Economy Act provides a legal basis to mandate some limited sharing with the Office for National Statistics). A mixture of regulation and institutional innovation is likely to be needed to enable greater provision of data by the private sector - discussed further below.

There is significant potential for shared or open data to promote competition and innovation in the economy, in contrast to the hoarding of data by digital companies with considerable market power, as recently noted by the Digital Competition Expert Panel.[42]

Institutions for the data economy

There is a vast literature on the appropriate institutional framework for provision of non-rival goods: what norms, regulations, and laws, and what mix of market, collective and government decisions about production and allocation will maximise social welfare? These questions are highly relevant to policy choices aiming to get the best out of the data economy. The amount of data is rapidly growing as digitization makes it possible to turn many goods and services into data records, and as behaviour is changing significantly shifting activities online. There are very many data sets, collected in different ways by

[39] 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/

[40] See for example https://blog.google/technology/research/open-source-and-open-data/

[41] Open Data Institute (2018) 'The role of data in AI business models', https://docs.google.com/document/d/14g0p6KSyH1r1J_PrykJIXUX-rdeP1B4CLIffAyFPOnk/edit#heading=h.rcydy9gttjg4

[42] 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.

different public and private sector organisations, with access restricted in varying degrees including by lack of interoperable technical standards. The number of potential uses probably far exceeds the actual usage of data to create valuable goods and services. The governance challenge is both to prevent misuse of sensitive data - where much of the public policy focus has been to date - and to realise more of the potential from data, ensuring the benefits are widely shared.

While this will involve traditional government regulation, discussed below, there is already some institutional innovation and experimentation with regard to access to data. Two principles are fundamental. First, in order to increase the economic value of data to society, the design challenges concern establishing terms of shared access enabling more use, capturing positive externalities while limiting negative ones.

Secondly, the trustworthiness of the institutions is of paramount importance as they will be determining who can access what data in accordance with the social and legal 'permissions' given. As O'Neill has argued, the real or perceived crisis of trust in many societies reflects suspicion of authority.[43] In the case of data, the suspicion can seem well warranted by frequent security breaches, stories of manipulation, or abuses such as Facebook/Cambridge Analytica. Informed consent, especially consent given to long and obscure terms and conditions online, is inadequate as a basis of trust. Instead, trustworthy institutions subject to intelligent forms of accountability (rather than the target-based or tick box versions found in some institutions) are needed. As Benedict Evans has pointed out, it is possible to discern an emerging societal consensus about who should be able to do what with different data: "Different entities have permission for different things."[44] Is it the supermarket, a video streaming app, or the police? Do we trust an organisation with certain data only as long as it is not too easy for them to use at speed or at scale, or too easy to join up with other data?

Although data is in its economic characteristics almost the opposite of a 'commons' (which refers to resources such as fish or grazing land that are rival in consumption), Elinor Ostrom's framework for the management of shared resources also offers some useful insights for data regulation and governance.[45] Her work considered contexts where people need to reach agreement about rules of access to a resource when some individuals will have to sacrifice private benefit for the greater common good. Just as a farmer upstream could benefit from not sharing water for irrigation with those downstream but will enable higher crop yields as a whole if they do participate, the holder of data may sacrifice some private economic rents by sharing but will unlock potentially much larger benefits for others.

She identified the conditions determining the way different goods are produced and allocated, including - as well as the characteristics of the good itself - the prevailing social norms and trust, the costs and benefits of different outcomes for different people, the information available and the technical or practical conditions. She also established the

[43] O'Neill, Onora (2002), 'A Question of Trust', www.bbc.co.uk/radio4/reith2002/
[44] Benedict Evans, 'Face Recognition and AI', https://www.ben-evans.com/benedictevans/2019/9/6/face-recognition
[45] Ostrom, Elinor (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

general design principles for collective self-managing institutions - set out here with implications for the data economy. Within a data economy we need to consider three groups of people or organisations involved in collaboration around data: those organisations which collect and share data (stewards); those that use data from those organisations; and those who are the subjects of data (ie the data is about them or is used in ways that affect them). These groups have different rights and capabilities, and there will be asymmetries of information between them. The table sets out Ostrom's design principles and their data economy parallels:

There are clear boundaries and rules about who is entitled to what	Requires debate about 'permissions' - which entities can access which data?
Monitoring actions is feasible	Requires transparency about terms and conditions, and auditability
There are mechanisms for resolving conflicts	These could range from withdrawal of access permission up to legislated penalties
Individual responsibilities and benefits broadly balance	Requires transparency and better understanding of value exchanges that are occurring, including short term gain for long-term cost
Users themselves are responsible for monitoring and enforcement	A minimum requirement is transparency and contractual terms that enable monitoring and auditing of all subsequent data uses/transactions; may involve agents acting on behalf of data subjects
Sanctions for abuse are possible and graduated, getting progressively tougher	Suggests enhancing current approach - more enforcement
Decisions are legitimated by the participation of users	For individuals, opt outs need to be viable; importance of competition so users have alternatives; trustworthy institutions have representative governance bodies
Decisions are also legitimated by government recognition	Comprehensive data strategy and legal/regulatory framework will be needed

These principles are useful for assessing the new types of institution or regulatory framework that will be needed to govern access to and use of data. They speak to the asymmetries of information and incomplete contracts characterising the data economy. Economic regulation in other domains is built on the extensive institutional economics literature, and the same analytical tools need to be brought to bear here in designing data access regulation - including the mandatory data access schemes under consideration in some jurisdictions as well as voluntary sharing arrangements..

Data Infrastructure

These institutional and regulatory questions need to encompass the whole of data infrastructure. Data infrastructure consists of:

- data assets such as datasets, identifiers and registers
- the standards and technologies used to curate and provide access to those data assets
- the guidance and policies that inform the use and management of data assets and the data infrastructure itself
- the organisations that govern the data infrastructure
- the communities involved in contributing to or maintaining it, and those who are affected by decisions that are made using it.

Schemes for sharing data previously kept closed for commercial or sensitivity reasons are starting to become more common, aiming to create a *club good* for the parties involved.

However, data sharing may be limited by a coordination problem. '[D]ata producers only have an incentive to make data available if they think there are enough users, and users need available data to get value from it. But data producers don't know how many potential users there are, and users don't know the amount, variety or quality of data that is available. This mutual uncertainty impedes data sharing.'[46]

Organisations considering sharing data with others face a number of other considerations:

- if the data in question contains personal information, there is a risk that a partner sharing it will mistakenly disclose it and incur regulatory and reputational costs for all involved;
- the data could reveal insights into the workings of the firm and its intellectual property;
- if the future use of the data being shared is unpredictable, whether because of the information it holds or as a result of its use being subject to novel technology, it is hard to determine whether a partner will invest sufficiently before the fact, or exit the arrangement at an undesirable time later on.

These questions are examples of the classic problems of asymmetric or incomplete information, principal-agent misalignments, and the difficulty of designing contracts under uncertainty.

Despite the barriers, a number of initiatives sharing data are under way. These schemes have the potential to change the incentives for governing data access. Schemes for shared

[46] 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', https://theodi.org/wp-content/uploads/2019/04/Datatrusts-economicfunction.pdf

data in a bounded space may change the returns on investment in the collection and cleaning of data, complementary skills and assets.

Data Trusts

One such approach to forming a data 'club', data trusts, are being developed and trialled in several countries. Schemes such as data trusts involve making more complete the contracts between parties that have asymmetric data holdings or technology skills.

A data trust can take a number of different forms - such as the legal trust, contractual, corporate, public, and community trust models. Arguments can be made for a plurality of approaches.[47] They have a number of aims in common:

- To enable data to be shared;
- To deliver public benefits as well as benefit of those sharing the data;
- To respect the interests of those with legal rights in the data;
- And ensuring the data is used ethically and in accordance with the rules established by the data trust;
- Ensuring that whoever holds data subject to the trust rules does so safely and securely, and that data is dealt with appropriately (for example by deletion) if the data trust ends;
- To manage individual rights and interests collectively (including any sharing of benefits received by the data trust);
- To set standard rules to govern all data sharing;
- To act as custodian/steward making decisions on behalf of data providers/ data users;
- And to be able to evolve to have new purposes, governance and working methods.

Trustees of a data trust may need to have powers strong enough to discourage misuse of the data, in line with Ostrom's principles.[48] Data trusts may be able to reduce transaction costs and increase efficiency, by allowing one data sharing agreement between partners rather than their having to negotiate several. They may be able to set conditions for the quality of data provided by members, perhaps reducing information asymmetries.[49] Data trusts may also be a way to compensate for 'missing markets'.[50]

The Open Data Institute has piloted data trusts based on contractual relationships between parties for sharing energy and mobility data in London, data about the illegal wildlife trade,

[47] Sylvie Delacroix and Neil Lawrence, 'Bottom up Data Trusts: Disturbing the 'One Size Fits All' Approach to Data Governance', forthcoming in International Data Privacy Law: Doi.org/10.1093/idpl/ipz014

[48] Register Dynamics (2019) "Putting the Trust in Data Trusts", https://www.register-dynamics.co.uk/data-trusts/

[49] Pinsent Masons (2019) 'Data trusts: legal and governance considerations', https://theodi.org/wp-content/uploads/2019/04/General-legal-report-on-data-trust.pdf

[50] 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', https://theodi.org/wp-content/uploads/2019/04/Datatrusts-economicfunction.pdf

and data about food waste; while Sidewalk Labs has used a data trust in its approach to the collection and use of data in an area of Toronto.[51]

Other data sharing models

Other approaches have also been adopted, either directly sharing datasets, pooling data through portals, or establishing platforms as mediators between providers and users of data.

One recent example is Databox, a multi-partner project funded by the EPSRC. It gives individuals control over the data they provide, including data increasingly being generated by Internet of Things devices such as smart thermostats and meters.[52] The data is held in a physical device controlled by the individual, rather than in the cloud, using 'containerisation' technology. According to a Royal Academy of Engineering Report, "Consumers will be able to obtain insights from their own data, while commercial organisations will have access to a greater range of data sources of appropriate type or granularity, enabling richer and more accurate analytics."[53] The Databox mediates access to the source of data but does not hold it. Individuals can give permission to third party app developers to access specific data. When the developer has used the data, the service can be provided to the individual without continuing to store data.

Data sharing in the UK energy industry has been mandated by the Government as part of the roll-out of smart meters. The Data Communications Company manages the smart meter infrastructure including data, licensed by Ofgem.[54] The in-home meter is linked to the telecommunications network enabling consumer data to be shared with competing energy suppliers, energy network operators and other authorised parties, such as third party intermediaries that offer energy saving, switching or load shifting services. Consumers are asked to authorise the use of their data. The infrastructure could potentially be extended: "Smart metering equipment could potentially be used to collect property information, such as temperature or humidity measurements, to spot where there are health risks to vulnerable people."[55]

Another example of data sharing by private companies required by government followed legislation (the 2017 Bus Services Act) mandating bus operators to share information. The

[51] See Open Data Institute (2019) 'Greater London Authority and Royal Borough of Greenwich pilot: What happened when we applied a data trust', https://theodi.org/?post_type=article&p=7891; Open Data Institute (2019) 'Illegal wildlife trade pilot: What happened when we applied a data trust', https://theodi.org/?post_type=article&p=7890; Open Data Institute (2019) 'Food waste pilot: What happened when we applied a data trust', https://theodi.org/?post_type=article&p=7889; Sidewalk Labs (2018) 'An Update on Data Governance for Sidewalk Toronto', https://www.sidewalklabs.com/blog/an-update-on-data-governance-for-sidewalk-toronto/

[52] Databox Project (2019) "Introducing BBC Box", https://www.databoxproject.uk/

[53] Royal Academy of Engineering, "Databox: allowing individuals to control how they share data with other parties", http://reports.raeng.org.uk/datasharing/case-study-1-databox/

[54] Data Communications Company, "What we do", https://www.smartdcc.co.uk/

[55] Royal Academy of Engineering, "Smart Meters: Data Sharing in the Energy Industry",http://reports.raeng.org.uk/datasharing/case-study-7-smart-meters/

Department for Transport created the Bus Open Data Portal and established standards and formats.

Instances of existing private data sharing models not mandated by legal or regulatory compliance include a DAFNI, a database and model repository for infrastructure providers; examples of 'open innovation' platforms such as APROCONE in aerospace or Goldcorp's then-startling (in 2000) opening of its proprietary geological database to invite outsiders to help locate gold deposits; and Strava Metro, which provides GPS tracking data from the Strava fitness app free to individuals and under licence to other users. In these examples, the incentives for data sharing vary, but there are clear benefits in each case to the companies sharing data: respectively, lower cost monitoring and enhanced resilience of infrastructure assets, design improvements along the supply chain, access to problem-solving resources, and building a reputation and customer base.

An alternative approach is Tim Berners-Lee's initiative Solid,[56] which centres on individuals controlling their own data, including terms of access and storage, in a decentralized model, in other words not involving any centralizing institutions. Users store data in one or more 'pods' (personal online data stores) hosted by an entity they can select, and they can permit different organisations to access data of different types. Solid's focus is therefore on individuals owning data they generate, and on safeguarding privacy. In other words, it is concerned with reducing negative data externalities from loss of privacy; to capture the potential social value from realising positive externalities, services and apps using data need to accumulate access permissions from individual users.

Although experience over time of using models of sharing may enhance trust and encourage growing participation, many shared data spaces - including most of the examples above - have required regulatory intervention. If the benefits of sharing are asymmetric, if the costs of building and maintaining a pool or platform are high, if there are concerns about loss of competitive advantage, or fears of regulatory or legal breaches due to handling sensitive individual data, a policy intervention will be required. Enabling the creation and capture of value from data, from new business opportunities and economic growth to improvements in public services and non-market gains, will require new policy approaches.

BOX: Transport

The transport sector in the UK illustrates a range of the issues discussed here.

Some public transportation and geospatial data is open and free. For instance, Highways England makes important data freely available to developers via an API.[57] The Geospatial Commission has launched a Single Data Exploration Licence (although users may need to purchase some of the data they identify).[58]

[56] https://solid.mit.edu/

[57] http://webtris.highwaysengland.co.uk/api/swagger/ui/index

[58] https://www.gov.uk/government/news/geospatial-commission-making-geospatial-data-more-accessible

With the public bus system, it has been a question of enforcing sharing of information. In the early 2000s, the Department of Transport (DfT) required the use of company data to inform public transit systems, improving services by providing users with more up to date information on buses' timetables, routes and fares. Bus companies had to share access to their real time operations, which also could allow for more effective monitoring of their performance. Transport Direct was set up as a distinct entity used by DfT to implement the opening up of company databases. Opening the data also created opportunities for other companies to create interfaces to inform the public about their transport options in real time.

Data does not have to be shared to improve outcomes for the public. For example, in response to a daily congestion charge on private hire vehicles in central London, Uber introduced a model in April 2019 that automatically adds £1 to every trip that passes through the congestion charge zone, regardless of time of day. At Heathrow airport, airport management has discussed the use of geo-fencing to to regulate private vehicle use around Heathrow.[59] Here, Uber has agreed to place limits on drivers to avoid congestion around Heathrow itself, using its internal dynamic pricing algorithm. In these two instances, Uber utilises its ability to adjust demand by altering the cost to the user, responding to the user's preferences. The end outcome on emissions and traffic is achieved through efforts within a private company rather than public sector regulation.

However, this latter case shows the capture of value from the use of data which is not shared. With the London congestion charge, depending on the number of rides and timing, Uber can collect the difference between its internal £1 congestion charge per ride and TfL's £11.50 charge per day. The distribution of revenue between the private company and public sector is known to Uber, but not necessarily to TfL. Similarly, while Uber is helping Heathrow Airport manage congestion, it alone is able to in effect implement a private congestion charge borne by drivers and passengers.

Yet private companies rely on the public sector to maintain the roads and public transport. The public sector remains responsible for the base map and road infrastructure. Importantly, the ordnance survey in the UK owns the coordinates system upon which transport services map their activities. Licences are required to use this base map, enabling information on the location of buses and so forth can be plotted. Nonetheless, the distribution of value, as private firms use data to manage transport services or traffic, is not necessarily equally, or at all, shared with the public sector. Service improvements through the use of data require negotiation not only about how information is shared but also how revenue is captured and distributed.

Issues around context, value capture and sharing data will become more pressing as the use of data for transport evolves. Autonomous vehicles illustrate the point made earlier that use affects the marginal cost and benefit of collecting more data. Autonomous vehicles require a classification model to identify and respond to different objects; the amount of data required for this model will reach a point at which diminishing returns set in. On the other hand, autonomous vehicles also require a base map of the world requiring ever more accurate and detailed information. Second, they show that sharing all data might not be needed to create value. Rather, running autonomous vehicles depends on

[59] https://mediacentre.heathrow.com/pressrelease/details/81/Corporate-operational-24/8878; https://www.thetimes.co.uk/article/heathrow-crackdown-to-beat-minicab-congestion-fhzgfzlqk

access to specific data at the moment when it is needed. Third, they also reveal some of the challenges around interdependencies and the distribution of value. Autonomous vehicles will rely on the base map and on road networks maintained by the public sector. Regulation - to ensure that some of the cost of providing this part of the data infrastructure is recouped by the public sector - will affect the market price of use of autonomous vehicles.

Box - health

Policy issues and recommendations

There are substantial barriers to the increased provision of shared data. These include the challenge of funding public goods with their cost structure of high initial but low marginal costs, and the trade off between wide availability of data and incentives to invest in its creation and provision, in both public and private sector. Furthermore, the benefits created by additional provision and sharing may be asymmetric, or costs may be imposed on the data holder in terms of loss of commercial advantage or additional risks. There are also significant concerns about privacy. Finally, regulation and the design of an appropriate institutional framework needs to address significant asymmetries of information and principal-agent problems.

Yet the potential economic benefits to society as a whole - not just a handful of commercial firms - of further data sharing and use are large. The basic economic principles point to the scope for gains from additional data provision and sharing if privacy concerns can be overcome, and a trustworthy institutional and regulatory framework established. The possibility of demonstrable widely-shared gains will be a precondition for trustworthiness.

We have analysed the social welfare value of data in terms of two lenses: its basic economic characteristics and its contextual, informational content:

Economic lens	**Information lens**
Non rivalry	Subject
Externalities + and -	Generality
Increasing/decreasing returns	Temporal characteristics
Option value	Quality
High fixed, low marginal costs	Sensitivity
Complementary investments	Interoperability/linkability

Our analysis of the actual and potential social welfare - society-wide economic value - in the data economy underlines the following principles:

- Market transactions alone will not bring about the maximum social welfare from data, given its economic characteristics of (positive and negative) externalities and non-rivalry;

- A more fruitful framing of the policy debate in order to generate increased social welfare from data, fairly shared, will be in terms of access rights and privacy protection, rather than ownership of personal data;

- Appropriate institutional and regulatory structures will be vital for a thriving data economy, regulating the permissions different types of entity have to access different types of data and monitoring and enforcing compliance. Work on the principles and structures of data governance for the maximum social welfare is in its early days and much more thought needs to be given to the specifics of regulatory and institutional design;

- New, trustworthy institutional structures are needed to develop to enable access to data in ways that make possible the creation of both commercial and wider social value, building on a range of approaches and pilots currently under way;

- Policymakers should recognise that the legal and regulatory framework they establish will affect both market and non-market values of data - the value of data is endogenous to the institutional framework;

- Additional approaches to quantified economic valuations, incorporating social welfare beyond private, market-based valuations, have limitations but will help improve understanding of the transactions taking place, particularly involving publicly-held data transactions with commercial organisations. In domains such as transport and health there is currently no public confidence that the terms of the deals will benefit the public. In addition to greater transparency, better understanding of data value is necessary;

- There are significant policy trade-offs including: between creating adequate incentives to invest in creating and maintaining data and related services on the one hand and the social value of widely diffused use on the other; and for public bodies between short-term financial gain from selling exclusive data access to the private sector and long-term economic and social gain from more open access;

- Contracts for data use are incomplete, and the regulatory framework should recognise this, particularly that schemes for sharing data in a regulated way change the returns on investment in collecting and cleaning data, and in complementary skills

and assets. The institutional and regulatory economics literature has many potential lessons for data regulation.

The detailed work required to flesh out these principles is beyond the scope of this report. The table below sets out some of the policy detail needing to be addressed:

Trade-off between investment/innovation and open/shared data	Are there parallels with IP frameworks - patent pools - or is this too complex? Compulsory licensing or franchising? Co-production rights? Is legal title to 'personal data' sufficient for privacy or are there better ways to protect privacy? Lessons from regulatory economics literature.
Financing data provision	Business models in the private sector; commercial models in the public sector. What charging mechanisms incentivise provision and also maximise social welfare? Are co-operative models relevant?
Enabling competition & growth	Codes of conduct applied to APIs; common technical standards needed. What privately-held data sharing needs to be mandated?
Regulatory thickets	Clearer guidance on sharing sensitive data (by public and private sectors) - overcoming the fear of breach of GDPR, fines. Models for communicating data use and access rights eg is there a parallel with simplicity of Creative Commons licences?
Terms of trade in public sector deals	Should public agencies ever grant exclusive licences to data? Data sharing as a licensing requirement eg for ride shares, smart city data, autonomous vehicles. Time limited licences. Are there lessons from spectrum auctions? Greater transparency needed for trust.
Mandating data provision/sharing by the private or public sector?	When is this needed? To what extent is Open Banking a model - for big tech companies? For NHS? Should public sector reference data all be open?
Institutions	Good models/metaphors? Trusts, pools, platforms, pods. What regulation/legislation is needed to establish a trustworthy framework. What forms of accountability are needed in both public and private sectors?

This report has set out a framework for thinking about how to increase value - in the broad economic sense of social welfare - in the data economy. Social welfare will be maximised by the ability to use data involving positive externalities and new options (while minimising negative consequences with regard to privacy), or in other words by identifying the potential for joining up data, creating new uses. The two lenses described here - economic characteristics (pp 4-8) and information characteristics (pp 8-15) - help identify which types of data may prove most valuable - and also the potential risks.

One of the concerns about the data economy is that big incumbent companies might continue to capture as private profit a large proportion of the value being created. They certainly have the greatest capacity to undertake the investment and deploy the specialist skills needed. However, preventing them from using data to provide valued services would be counterproductive. A more effective way of bringing about a more even sharing of the economic welfare created by data use would be the direct approach of using competition policy to open the data-driven markets to other providers.

This requires policies addressing the challenges described in this report in a systematic way. Considerable work is needed to fill out the details of the framework set out here. Three avenues stand out. One is attempting quantification, as sketched out here, in some specific data domains; models from financial economics may be applicable. A second is translating the economic and information lenses into a practical toolkit, particularly for the use of public sector organisations. Finally, the challenges of regulatory and institutional design in a context of information asymmetries, principal-agent problems and pervasive externalities is a problem the body of work in institutional and regulatory economics ought to be able to address.

One final note is that the scope for use of data to increase social welfare - to provide better public and private services to everyone - demands a strategic policy approach.While many countries have adopted a range of open data policies, comprehensive strategies are rarer; the UK and Canada are developing National Data Strategies. The precautionary principle is often applied in the context of unknown future risks; in this context of unknown future opportunities perhaps an optionality principle should apply, a presumption in favour of creating the conditions for greater access, sharing and use of data, within a framework of appropriate trustworthy institutions and safeguards.

Advisory group members

Azeem Azhar, Exponential View
Joshua Ballantyne, DCMS
Claire Craig, Queens College, Oxford
Catherine Dennison, Nuffield Foundation
Ray Eitel-Porter, Accenture
Jonathan Haskel, Bank of England MPC
Richard Heys, ONS
Herman Heyns, Anmut
Ed Humpherson, Office of Statistics Regulation
Frank Kelly, University of Cambridge
David Knight, DCMS
Rannia Leontaridi, BEIS
Wendy Li, US Bureau of Economic Analysis
Stephen Lorimer, Greater London Authority
Sergi Martorelli, Glass AI
David Nguyen, NIESR
Reema Patel, Ada Lovelace Institute
Charles Price, HM Treasury
Marshall Reinsdorf, IMF
Chris Riley, Mozilla Foundation
Eric Salem, Office for Life Sciences

Interviewees

Andrew Dilnot, Nuffield College, University of Oxford
Herman Heyns, Anmut
Frank Kelly, University of Cambridge
Derek McAuley, University of Nottingham
Sergi Martorell, Glass AI
Richard Mortier, University of Cambridge
John Taysom, Privitar
+ 7 others - permission to be checked

We are grateful to the Nuffield Foundation for funding this project, Valuing Data: Foundations for Data Policy under grant number WEL/43956.

The Nuffield Foundation is an endowed charitable trust that aims to improve social well-being in the widest sense. It funds research and innovation in education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation. More information is available at www.nuffieldfoundation.org.