European Commission

JRC TECHNICAL REPORTS

The semicircular flow of the data economy

Pablo de Pedraza
Ian Vollbracht

2019

Joint Research Centre

EUR 29825 EN

Contact information
Name: Pablo de Pedraza
Address: European Commission
DG Joint Research Centre
Directorate I — Competences
Unit I.1 — Monitoring, Indicators and Impact Evaluation.
Via E. Fermi 2749, I-21027, Ispra (Va) - Italy
Tel. +39 0332 78 3805
E-mail: pablo.depedraza@ec.europa.eu; JRC-COIN@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC117362

EUR 29825 EN

PDF	ISBN 978-92-76-09231-5	ISSN 1831-9424	doi:10.2760/668
Print	ISBN 978-92-76-09232-2	ISSN 1018-5593	doi:10.2760/40733

Luxembourg: Publications Office of the European Union, 2019

How to cite this report: Pedraza Garcia, P. de and Vollbracht, I., *The semicircular flow of the data economy*, EUR 29825 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-09231-5, doi:10.2760/668, JRC117362.

Contents

Acknowledgements

The authors would like to acknowledge comments and suggestions from Bertin Martens, Daniel Vertesy, Michaela Saisana, Sven Langedijk, and Ignacio Sanchez and previous work and discussions conducted at Webdatanet Task Force 25 during the cost action webdatanet (IS-1004, www.webdatanet.eu) and at a series of seminars held at the European Commission Joint Research Centre in Ispra, Italy. The scientific output expressed does not imply a European Commission policy position. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

Authors

Pablo de Pedraza

Ian Vollbracht

Abstract

This paper revisits the traditional 'circular flow' of the macroeconomy (Samuelson, 1948) and reworks it to capture the use of big data and artificial intelligence in the economy. The characterisation builds on the multifaceted role of data to conceptualise markets and differentiate them depending on whether data is an output, a means of payment, or an input in knowledge extraction processes. After this, the main differences between the circular flow economy and the data economy are described, identifying the new flows and agents and the circular flow assumptions that do not seem to be as relevant to the workings of the data economy. The result is a 'semicircular' flow diagram: unprocessed data flow from individuals, families, and firms to data holders. Only data processed in the form of digital services flows back to families and firms. The new model is used to explore the potential for market failures. Knowledge extraction to generate digital services occurs within a 'black box' that displays natural monopoly characteristics. Data holders operate simultaneously in the markets for data generation and knowledge extraction. They generate the amount of knowledge that maximises their profit. This creates data underutilisation and asymmetries between data holders and other agents in the economy such as anti-trust authorities, central banks, scientific communities, consumers, and firms. Public intervention should facilitate additional generation of knowledge by developing additional merit and non-rival uses of data in such a way that knowledge generation maximises the social gain from digitalisation. The semicircular model can incorporate data leakages and knowledge injections activated by data taxation. Data taxes should be paid with data respecting existing legislation, privacy concerns, and preserve the incentives of the data holder to innovate in competitive data generation markets. A centralised data authority, as initially proposed by Martens (2016) and more recently by Scott Morton et al. (2019), would be responsible for knowledge generation and aim to achieve better regulation, standards, and transparency, and maximise common good. Our conclusions are in line with an extensive user-centric approach to data portability (De Hert et al., 2018). This paper contributes to the digital economy discussion by developing a simple theoretical motivation for increased access to data for the public good, which will stimulate further theoretical and empirical exercises and lead to policy actions.

Keywords: big data, artificial intelligence, macroeconomy

1. Introduction

For several years many journalists and researchers have claimed that the drastic reduction in the cost of collecting, storing, analysing, and obtaining valuable knowledge from diverse data sources has changed the way the economy works. In this view, markets are more and more driven by big data (BD) and artificial intelligence (AI) and by their interaction. Although there is an increasing number of conceptual and normative analyses of the digital economy (Codagnone and Martens, 2016), the standard circular flow model of the macroeconomy that conditions the way that many policymakers think about the economy at large has no explicit role for data (Duch-Brown et al., 2017). As far as we know there have been no attempts to revamp the circular flow model of the economy to take the role data plays in the economy into account.

Some authors and thinkers have gone so far as to suggest that a world of advanced AI and BD techniques will lead to the 'end of theory' (Anderson, 2008; Kitchin, 2014, Couper 2013) as all 'solutions' can be analysed from a reductive process of feeding huge data sets into self-optimising but 'theory-free' AI algorithms (Silver et al., 2017). AI and BD are certainly already challenging epistemologies across the social sciences and seem set to do so to an ever greater extent in the future. However, for as long as economic policymaking continues to pass through human-led decision-making processes and 'filters', economic theory will continue to provide a powerful conceptual framework for thinking through policymaking and regulatory decisions. In this context, the circular flow of the economy is the simplest macroeconomic model. It has traditionally been used to teach how the market economy determines optimal production (Samuelsson, 1948; Samuelsson and Nordhaus, 2010) through the interaction of households, firms, and governments in factor and product markets. It helps students to understand the functioning of the 'invisible hand' (Smith, 1778) and the failures that justify government intervention in the economy. This paper considers how the traditional circular flow model of the economy might be adapted to the 'data economy' by examining the extent to which the main underlying assumptions continue to hold in a data economy in which BD and AI techniques have superseded the 'pin factory' (Smith, 1778) model of production. Both similarities and differences are therefore identified to develop a new version to illustrate potential market failures and how government intervention might evolve in a BD- and AI-driven data economy.

The new model was constructed by taking the following differences into account.

Firstly, traditional economy flows of services, tangible goods, factors of production, and money can be synthesised in terms of factor and product markets. Incorporating BD flows and the use of AI to extract knowledge from data blurs this distinction. The main reason is that data plays a **multifaceted role** in the economy. Sometimes data is an implicit means of payment that consumers use to pay for digital services (Facebook, for example). Sometimes data is an explicit output of consumption processes that data holders collect (such as credit card spending data),and sometimes data is a factor of production in itself, acting as an input for the knowledge extraction process out of which digital services are created (such as marketing companies targeting consumers or insurance companies using BD and AI techniques to price their policies).

Secondly, although data is a means of payment, 'atomistic' data points are **difficult to price** in isolation. The value of data is therefore unclear, which in turn makes the price formation process of payments in the data economy obscure. Indeed, one atomistic data point that an individual uses for a specific payment may, in isolation, have no value for BD and AI approaches. However, if that atomistic data point is included in a linked data set that covers a range of aspects of an individual's life, its value changes. Furthermore, if that set of individual data is part of a fully searchable and integrated data 'lake' covering a vast number of atomistic individuals, then the value is quite different again. This 'contextual' value of data is something that is not explicitly dealt with in the standard circular model, which assumes explicit price formation for inputs and outputs. In contrast, data pricing in the data economy is very often 'opaque' and 'contextual'.

Thirdly, only large data aggregators with large computing machines can capture the positive externalities from data aggregation. The production function of data collection and processing therefore displays **economies of scale and scope** (Scott Morton et al., 2019), setting data holders in a race towards N = All and X = Everything. The point at which diminishing returns to scale arrive in BD and AI knowledge generation processes is an unresolved empirical question, but if the marginal cost of an additional unit of data storage is assumed to be zero and the (potential) added value of an additional data point is (potentially) positive at the margin, then the **theoretical** implication is that the point of diminishing returns is never reached. In other words, the bigger the data lake, the more efficient the data processor in question is (Blake et al., 2014). Notably, if it is assumed for the sake of argument that AI and BD technology is common to all firms, then the size of the data lake held by a given firm is what constitutes its core competitive advantage. Network effects and sunk costs reinforce the natural monopoly characteristics of knowledge extraction.

Fourthly, the resulting flows of economic interactions are circular in the sense that atomistic individuals, such as citizens and firms, generate data (which is used by data holders) and receive services as part of the exchange. However, the model is **semicircular** in the sense that direct data flows of unprocessed data are unidirectional from citizens and firms to data holders but not from data holders to citizens, firms, or governments. Data flows from data holders to other agents in the economy include a wide range of taxonomies but are only indirect in the sense that they are processed in the form of customised goods or services (Bergemann and Bonatti, 2018; Bergemann et al., 2018). For example, consumers trade their primary personal data in return for a data-driven service on Google's search engine or on Facebook. What flows back to consumers is non-monetised consumer surplus but not data. There is no way to 'trace' the way in which BD and AI techniques make use of individual data points. This semicircularity and lack of traceability is an additional difference from the explicit and traceable flows of capital and labour in the standard circular flow approach. This makes the semicircular flow similar to creative processes that are elusive to patent and copyright regulations, thereby 'de facto' generating an insufficiently transparent situation.

Fifth, **data property rights** are often unclear in practice ([1]). From an economic point of view, data holders are 'de facto' owners of a great many large data sets. However, theoretical categorisation of the data as public, private, club, and/or merit goods is not clear and may depend on the specific use of the data in question. The current de facto ownership constrains alternative merit and non-rival uses of data that could increase knowledge and social well-being without damaging privacy and the interests of data holders. Optimum degrees and forms of data sharing or access are yet to be defined (OECD, 2015; Palfrey and Gasser, 2012; Scott Morton et al., 2019). Scale- and scope-driven centralisation and underutilisation generates information asymmetries between data holders and other agents in the economy such as consumers, other firms, regulatory authorities, and the scientific community. De facto ownership operates as a 'breastplate', a shell that allows generation of knowledge from data within an AI 'black box'. This only allows uses of data that maximise the data collection of data holders and their profit. It benefits society because it protects the investments of data holders and their incentives to keep on innovating so that AI expands to cover more human activities (towards X = All). It also bears an opportunity cost for the whole of society as additional alternative merit and non-rival uses of data could generate large social gains and further innovation.

The first difference is composed of the starting point for conceptualising markets in which data play a prominent role and differentiate between them according to the role that data plays. On the one hand, there are pure and mixed data production markets in which implicit prices are paid using data. Data in these markets is produced/generated as an explicit **output**. On the other hand, knowledge extraction markets are markets where data is an input. Markets where data play different roles do not necessarily have the same structure. Incorporating these new markets implies the incorporation of a new type of actor/stakeholder: data **holders**. In practice, data holders (European Commission, 2018a; European Union, 2016 ([2])) are very often digital giants such as Facebook or Google that operate businesses of this type on a vast scale. Data holders are the de facto owners of data (Duch-Brown et al., 2017). Data protection legislation is developing rapidly around the world. The EU's general data protection regulation (GDPR), (European Commission, 2016), which came into force in 2018 is a

([1]) The EU's general data protection regulation has greatly improved personal data protection and has certainly reduced levels of improper use of personal data. Nonetheless, within given purpose limitations, firms remain able to make profitable use of integrated data sets comprising data points that may atomistically be considered 'personal'.

([2]) The GDPR (European Commission, 2016) refers to 'data controllers' and 'data processors'. According to Art. 4, a 'controller' is a natural or legal person, public authority, agency, or other body that, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union law or the law in a Member State, the controller or the specific criteria for its nomination may be provided for by Union or Members State law. A 'processor' is a natural or legal person, public authority, agency, or other body that processes personal data on behalf of the controller. Many data processors are also data controllers. Data holders are referred to with the aim of conceptualising an economic agent, as explained in the text.

global benchmark. This paper assumes that — in practice — data holders exist and are able to amass data lakes of varying sizes, to which AI and BD techniques are then applied to provide value-added services to the economy. They operate as de facto owners of data (Duch-Brown et al., 2017) in a manner that should be fully compliant with existing regulations. However, their activities very often occur in a black box (European Commission, 2018a) which questions the need for clearer and more precise allocation of rights over data (Dosis and Sand-Zantman, 2018) and probably additional regulations (De Hert et al., 2018).

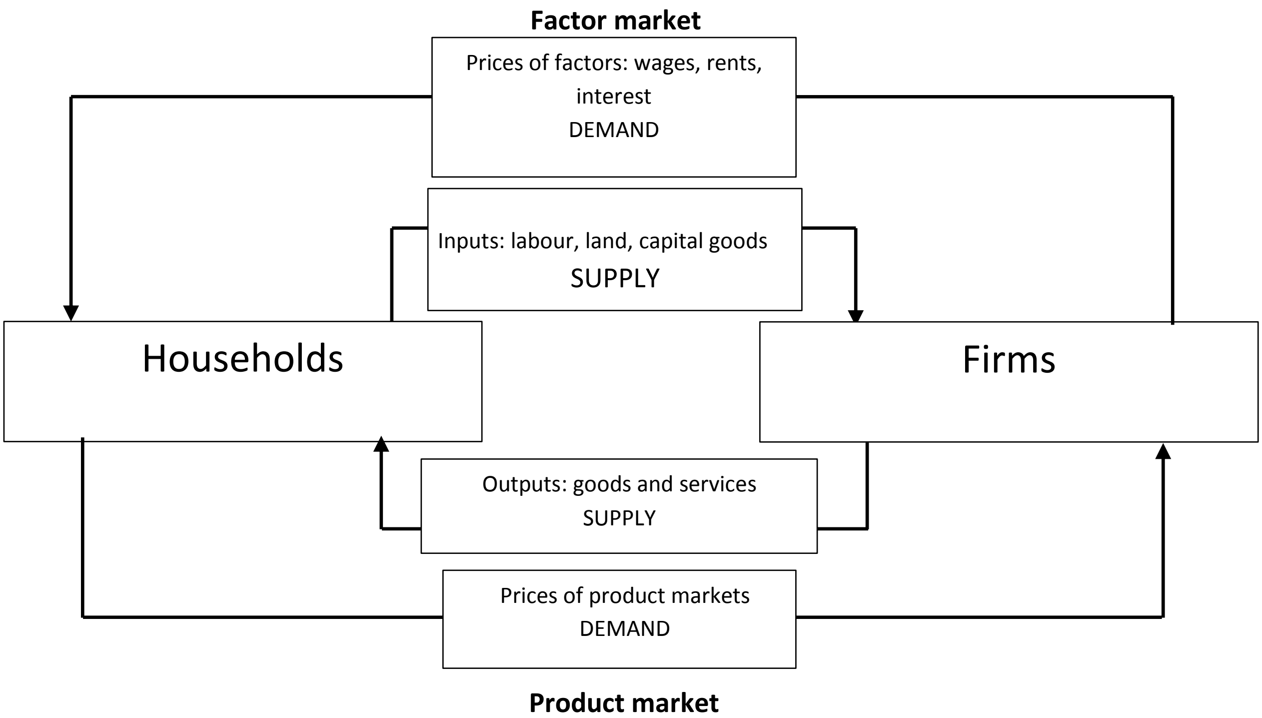
The semicircular model presented in this paper improves understanding of the interaction of households, firms, data holders, and governments in data production, knowledge extraction, factor and product markets, and the failures that justify government intervention in the economy, and to propose future research lines and policy actions required for optimal knowledge generation.

This paper is structured as follows. Section 2 describes the traditional circular flow of the economy model and its equilibrium. Section 3 develops the semicircular flow of the economy framework by defining data production markets and knowledge extraction markets and identifying the circular flow assumptions that fail in this new model. Section 4 goes deeper into the market failures that characterise knowledge extraction and the economic characterisation of data. Section 5 reviews the information asymmetries that knowledge extraction accompanied by de facto ownership generates, and which leads to the semicircular nature of the digital economy. Section 6 focuses on the policy actions motivated and stimulated by the semicircular characterisation of the data economy. Finally, Section 7 draws conclusions and suggests directions future research might take.

2. The circular flow of the economy

The diagram of the circular flow of the economy below represents the major exchanges in the economy (Samuelsson and Nordhaus, 2010). There are factor markets and product markets. In factor markets individuals provide labour (input) and, in exchange, businesses provide individuals with wages (income). Firms also spend on other factors of production such as capital and raw materials, and transfer income to the factor owners. In product markets individuals buy the goods and services that businesses produce. Factors of production enable businesses to produce goods and services (output) and income allows the owners of factors and workers to buy goods and services. Money is the means of payment that everybody accepts and is used to measure economic value.

Figure 1. The circular flow of the economy



In Figure 1, flows and their direction are represented by black arrows. The expenditure of sellers (employers/firms) becomes income for buyers (workers/households) and vice versa. In this simple, theoretical model, aggregate **expenditure** of an economy is identical to aggregate **income.** The economy is characterised by **specialisation** and use of capital, which make labour more productive and allow a network of trade to develop and extensive use of **money**. Economic activities are coordinated by the market without government intervention. In every market there are two sets of agents, buyers and sellers, who interact in exchanges which determine prices and quantities of goods, services, and assets. No overall organisation is responsible for production, consumption, distribution, or pricing. **Explicit prices** play a central role in balancing supply and demand in each individual factor and product market. They serve as signals to producers and consumers and coordinate their decisions.

Before any exchange, agents will incur search costs to find their match; there is **no centralised specific agent** responsible for matching efficiency that gathers information from both sides and centralises decisions about the best matches ([3]).

([3]) In most markets there is a place where agents meet: in housing markets through a real estate agent, in shares markets through the stock market, and those looking for a partner through dating agencies. Marketplaces speed up the rate at which traders find each other and match (Coles and Smith, 1996). In macroeconomic models the matching process is traditionally captured by a Cobb-Douglas matching function such as:

$$Log\ (M_s) = \log(\lambda_s) + \beta_1 log(B_s) + \beta_2 \log(S_s)$$

More complex circular models incorporate imperfections that exist in the real world which can lead to unemployment, financial crisis, monopolies, negative externalities such as pollution, or ethically unacceptable income distribution, and extreme poverty. Governments then have a role in seeking to 'correct' these failures by aiming to increase **efficiency**, promote **equity**, and foster macroeconomic **stability** and growth ([4]). Markets are not efficient when there is no perfect competition as in the case of monopolies, externalities, or public goods. Markets are efficient in the case of private goods. Monopolies lead to lower output, moving the economy inside the production possibility frontier and increasing prices, which motivates government intervention to break them up. Merit goods with positive externalities such as education are very often financed by governments. Whenever market outcomes result in an unfair distribution of wealth, the state may develop redistribution policies, and when the economic cycle goes through periods of low economic activity, the state may seek to stimulate the economy by means of fiscal and/or monetary policies.

Figure 2 augments the circular flow of the economy to include leakages and injections. Agents can withdraw or inject money from the circular flow of an economy. Leakages (red arrows) and injections (blue arrows) occur in the financial, government, and overseas sectors. Financial institutions or capital markets play the role of intermediaries providing the option to withdraw money from the flow by saving (S) money or to inject investment (I) by borrowing money. The financial system is also a circulatory system of funds in which, technically speaking, lending from savers is equal to the borrowing from investors. Financial assets and instruments are denominated in money. Central banks set interest rates that are the price of borrowing money. **Taxes** (T) are a leakage from households and firms to the government sector that makes government activities financially feasible. Government activities, subsidies, **welfare transfers** to the community, and purchases of goods and services provide injections (G). Government obtains revenues to promote efficiency, equality, stability, and law enforcement from taxes. Taxes are like the price of public goods with the difference that private consumption is voluntary while taxes are not. Governments exercise their monopoly of power to collect taxes. Flows of goods, services, and finance occur across national borders. Overseas economic relationships make it possible to import (M) from the rest of the world, which represents a leakage to other economies, and export (X), which injects income from overseas. International law and international institutions regulate international relationships such as trade.

The **state of (macro) economic equilibrium occurs** when total leakages (savings (S) + taxes (T) + imports (M)) are equal to the total injections (investment (I) + government spending (G) + exports (X)) that occur in the economy. This can be represented by:

$$S + T + M = I + G + X \quad (1)$$

Disequilibrium occurs when leakages are not equal to the total injections. In such a situation, changes in expenditure and output will lead the economy back to the state of equilibrium. Such changes will depend on the type of inequality:

---

where the number of successful matches (M_s) in markets is explained by the number of buyers (B_s) and sellers (S_s) in that market and λ_s, the efficiency of the matching process between them. When variables such as geographical location and other characteristics of goods and services match those demanded by buyers, λ is higher, and the matching process is more efficient. It is also determined by buyers' and sellers' search efforts that may be channelled by contacting specific agencies, but there is no centralised agent in charge of matching efficiency.

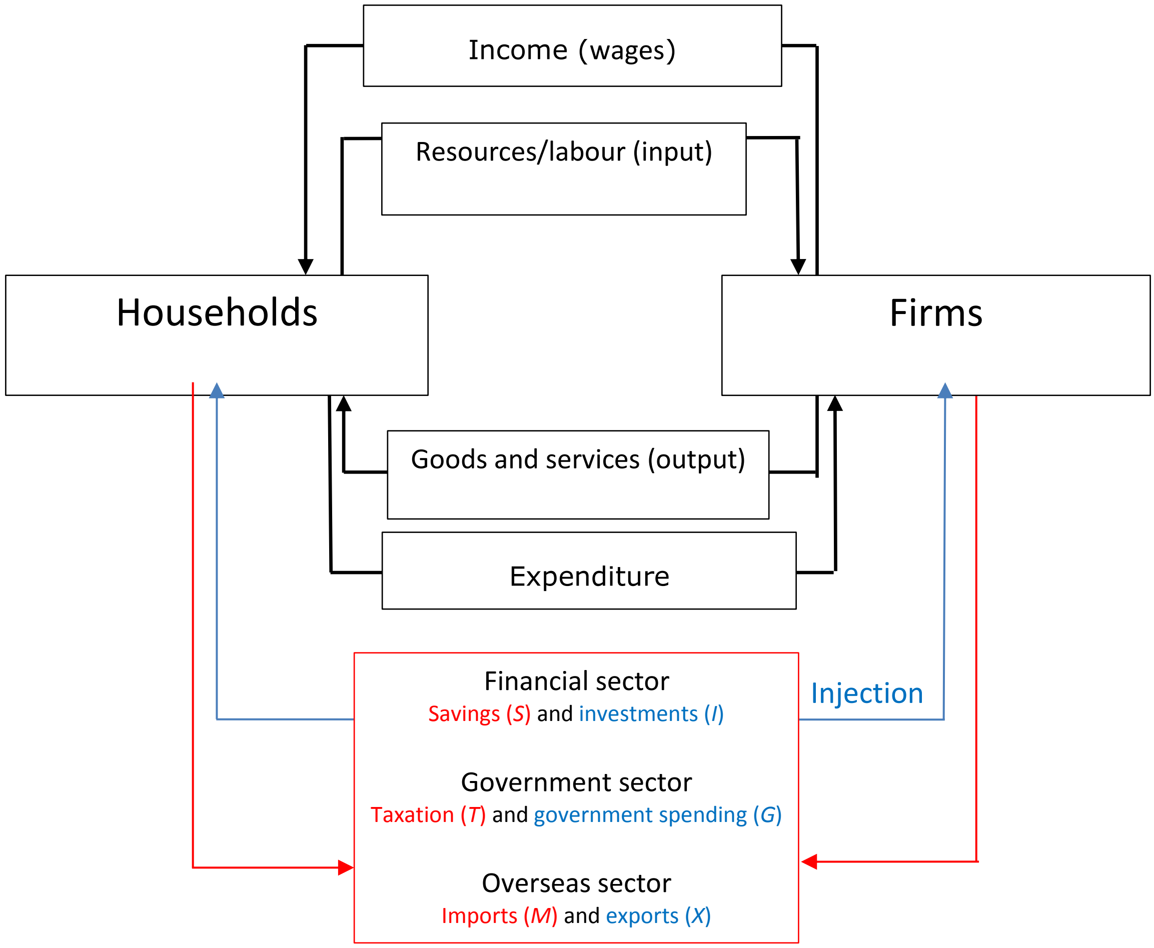
([4]) **Efficiency** of perfect competition implies that markets will produce the goods and services most desired by consumers, using the most efficient techniques and the minimum amount of inputs given the production possibility frontier (PPF). There is no perfect competition in the case of monopoly elements, externalities, or public goods. In the case of monopoly, a seller is able to affect the market price leading to higher prices and lower output moving the economy inside the PPF. Governments regulate monopolies and introduce anti-trust laws to prohibit actions such a price fixing or open markets to competitors. Externalities occur when cost or benefits are imposed on third parties not participating in the market. Government regulations are designed to control negative and promote positive externalities. Pollution is an example of a negative externality, while the spillover from research and development to the whole society is a positive one. Cases of positive externalities are public goods such as national defence, and merit goods such as education and health. They are often financed by governments because private incentives for their production do not always exist, and several problems may arise if they are not efficiently provided. Markets may function efficiently but produce an unfair distribution of income. Income inequality may be ethically unacceptable and voters may decide to change the government. Governments can reduce income **inequality** by progressive taxation, transfer payments, or by providing basic necessities such as education and health care. In addition, capitalism has always suffered periods of **instability**, fluctuations, business cycles, and periods of inflation or high unemployment. Governments can affect output, employment, and prices by means of fiscal and monetary policies and build a social safety net for the elderly, unemployed, and impoverished people. It is clear that an efficient human society requires both markets and governments.

$$S + T + M > I + G + X \quad (2)$$
or
$$S + T + M < I + G + X \quad (3)$$

In equation 2, if the sum of savings (*S*) plus taxes (*T*) and imports (*M*) is greater than the sum of investment (*I*) plus government spending (*G*) and exports (*X*), income, output, and expenditure will fall, contracting economic activity. Families' incomes will be lower, which will reduce their savings, the income they spend on imports, and the taxes they pay. This will lead to a reduction in the left-hand part of the equation until the economy is back in equilibrium (equation 1). Governments can accelerate the process to achieving equilibrium by increasing government expenditure. They can, for example, alleviate the reduction in family income by transferring rents through unemployment benefits or subsidies.

In equation 3, if the sum of savings (*S*) plus taxes (*T*) and imports (*M*) is smaller than the sum of investment (*I*) plus government spending (*G*) and exports (*X*), income, output, and expenditure will rise, expanding the economy. As families' incomes increase so do their savings, their expenditure on imports, and the taxes they pay, leading the economy back to equilibrium (equation 1). Equilibrium can also be explained from the micro-economic perspective as a result of consumers maximising their utility, producers maximising their profits, and a decentralised matching process.

Figure 2. The circular flow of the economy, leakages, and injections.



3. New markets, new flows, new agents, and new assumptions

The presence of BD and AI brings new flows, new markets, and new stakeholders and the need to revise some of the above assumptions and find a new way of representing the major exchanges in the economy.

Based on the role played by data, three new markets can be identified. On the one hand, **pure data production markets** and **mixed data production markets** are substantially different from traditional factor and product markets but increasingly interconnected and operating in both of them and affecting the entire economy (Arthur, 2011; Codagnone and Martens, 2016). Their main characteristic is that consumer activities in both markets generate BD. As a result, a new type of stakeholder emerges: **data holders** (European Commission, 2018a; European Union, 2016). Data holders are profit maximisation companies that specialise in collecting data in pure and mixed data production markets, obtaining de facto ownership (Duch-Brown et al., 2017), processing the data (Dosis and Sand-Zantman, 2018), and re-selling after processing in the form of customised goods or services (Bergemann and Bonatti, 2018; Bergemann et al., 2018) to either other firms or consumers. They include businesses such as social networks, search engines, network operating systems, e-commerce, or sharing (Scott Morton et al., 2019). Data holders accept data as an implicit payment for the provision of services in data generation markets where their activity is very often not profitable in monetary terms (Bond and Bullock, 2019; Kaminska, 2016; McArdle, 2019). They extract profitable knowledge and information from data by operating in the **BD knowledge extraction markets** where they can monetise BD (Scott Morton et al., 2019). In this paper the term data holders only refers to private companies. There are also huge data lakes in the public domain that may currently be underutilised such as tax records and the health data system, which also have the potential to be used as BD and AI data lakes for creating value-added analytical services. This section conceptualises and characterises pure and mixed data production markets and BD knowledge extraction markets and incorporates them in a new version of the circular flow of the economy: the **semicircular flow of the data economy** because of the one-directional flow of unprocessed data from households and firms to data holders. It reviews the assumptions of the circular flow of the economy, which fail in the resulting semicircular flow model.

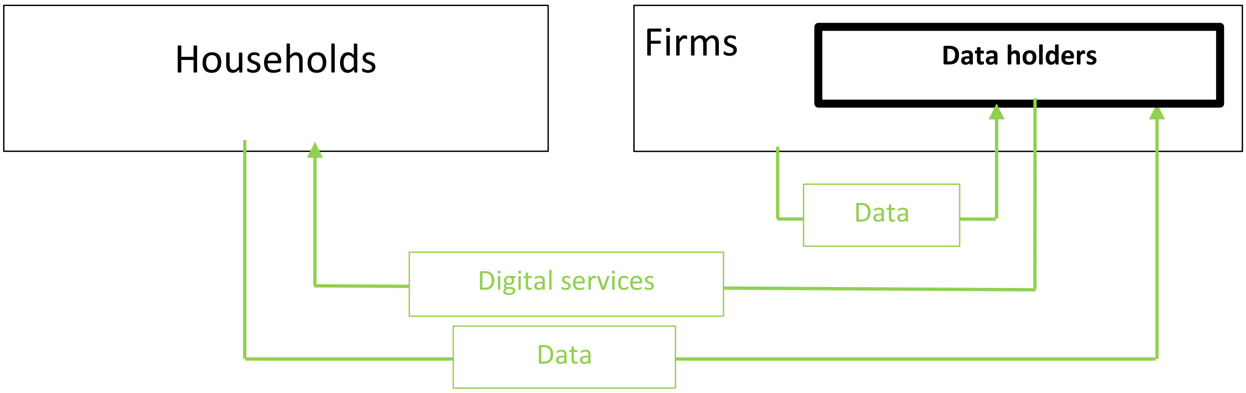
Pure and mixed data production markets

Pure data production markets are **barter markets** in which no explicit medium of exchange such as money is used (Evans, 2013; Scott Morton et al., 2019; Tett, 2018). Data holders provide 'digital services' that are the result of knowledge extracted from BD using AI techniques. Consumer payment in return for that service is made using the data and only the data that they generate while using those digital services. This is the case for search engines such as Google and social networks such as Facebook. This is also the case for LinkedIn, Dropbox, and Spotify in their 'freemium' basic versions (Kramer and Kalka, 2016) ([5]). Figure 3 shows these flows of exchange of digital services for data using green arrows. These companies are data holders that do not obtain a direct monetary compensation from this activity but an implicit price in data. There would be no incentive to invest in these services if it were not for two things: first, at least de facto data processor (service provider) ownership of data, and, second, another market where data holders can monetise the data (Dosis and Sand-Zantman, 2018; Jones and Tonetti, 2018, Scott Morton et al., 2019) and therefore obtain a **revenue stream** from them**.**

Non-monetary prices, or implicit data prices, change the theoretical representation of supply and demand in the price and quantity space and in the calculation of consumer surpluses (Brynjolfsson et al., 2018). Data does not have a clear value, especially for consumers who have very little idea of the data being collected (Scott Morton et al., 2019), which blurs and dilutes utility maximisation and the calculation of the demand graph and its slope.

([5]) For example, Facebook offers a 'free to use' digital service that allows people to stay in touch. Users generate data when they create a user profile indicating their name, occupation, schools attended. and when adding other users as 'friends', exchanging messages, statuses, pictures, videos, links, 'likes', and other Facebook reactions together with the exhausted data (paradata, environmental data, or footprints) related to their activity. Other examples are Instagram, a photograph- and video-sharing service, or WhatsApp, a messaging service, which are free to use and do not generate direct revenue but do generate data.

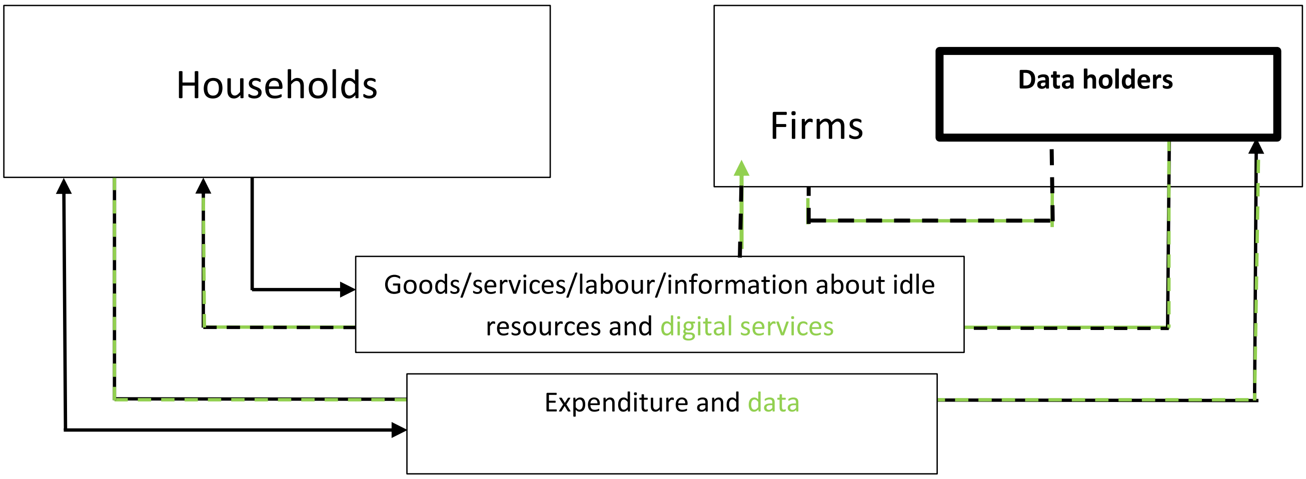
Figure 3. Pure data production markets and the circular flow of the economy



Mixed data production markets are those in which data holders obtain two rewards in return for their service. Firstly, as in pure data production markets, they collect the data generated by users. Secondly, they receive payment in the traditional way as an additional compensation. This is the case of LinkedIn, Dropbox, and Spotify in their 'premium' upgraded versions (Kramer and Kalka, 2016). This is also the case of multi-sided markets and platforms in the sharing economy such as Uber or Airbnb in which consumers pay a service fee. In these platforms, data holders offer search cost reduction services facilitating supply and demand matching after extracting knowledge and information from data.

Figure 4 represents these flows of exchanges using dotted green and black arrows. The arrows are green because there is a two-directional exchange similar to that of pure data production markets: citizens receive a digital service and, while using it, they pay part of the price by generating valuable data about their experiences and behaviour. The black arrows indicate that, as in the traditional product markets, there is a flow of offline goods and/or digital services in one direction and **monetary expenditure** in return for it in the other direction.

Figure 4. Mixed data production markets and the circular flow of the economy



Many online platforms facilitate matching between households that are willing to share their assets such as houses or cars while idle or while not being fully used (Airbnb, Couchsurfing, Zipcar, Uber, Lyft, BlaBlaCar, Turo — formerly RelayRides — Getaround, etc.), to recirculate goods (eBay, Etsy, Freecycle, Freegive, Yerdle, SwapStyle, etc.) or even offer services and labour (TaskRabbit, myTaskAngel, Freelancers, etc.) or just to build social connections (Codagnone and Martens, 2016).

Processing of BD by means of AI facilitates matching between households. This 'sharing' of goods and services for a monetary payment or for free, and flows of goods and monetary payments from household to household, is represented by two black arrows. The arrows are black because, although the exchange takes place between families (consumer to consumer, or C2C), it represents the same flow of goods and services that happens in traditional markets. There is not a green flow of data from household to household because participants receive information and knowledge on only a limited set of offers that result from search rankings or matching algorithms. Although there are no data flows from household to household, thanks to reductions in search costs, households can supply goods and services previously provided by **private specialised firms**. These peer to peer (P2P) or sharing flows also occur from firm to firm. Similarly, **search frictions between firms** (business to business, or B2B) and between firms and consumers (business to consumer, or B2C) can be reduced to benefit both sides of the market or the interests of data holders.

Data flow in pure and mixed data production markets is one directional, semicircular, and from households and firms to data holders. Both pure and mixed data production markets are at the same time **consumption markets** and **production processes**. On the one hand, there is an exchange and, on the other, there is a data production process about consumer (and firm) activity and behaviour. Both markets have a data flow from households (and firms) to data holders and a flow of 'algorithmic' services such as matching algorithms or search rankings from data holders to households and firms. Both markets are data holder data factories in which they **produce feedstocks** that are used to improve existing services and AI and to develop and power other services. More and more devices contain sensors and more activities operate in the same way as these two markets, generating data as part of the payment. The economy is characterised by increasing capacity to pump zettabytes of unstructured data to data holders (Economist, 2017), who as owners of data factories, obtain de facto ownership of data.

Big data knowledge extraction markets

Data holders accept payment in data because they are able to generate revenue streams from them in the **BD knowledge extraction markets**. Data in these markets is processed, refined, valued, purchased, sold, exchanged, or merged for the purpose of maximising revenue streams, generating and collecting additional data, and improving AI. Knowledge extraction markets operate as if in a **black box** in which BD is the feedstock of a process producing knowledge, information, insights, patterns, predictions, and new services and AI improvement. Search engines and social networks such as Facebook and Google operate in pure and mixed data production markets but base most of their revenues on advertising and operating as B2C and in B2B intermediaries. Market platforms also operate as C2C or B2B data intermediaries. They use personal data to communicate with their customers and provide better and more personalised services, make suggestions and advertise to market their own products or offer marketing services to other firms. They also use data to train and improve algorithms, that is, the more people use a search engine, the better it becomes which allows it to generate new AI services such as image and face recognition, translation, or personality assessments through social networks. Knowledge extraction from BD is influencing an increasing proportion of the economy. Firms from diverse industries are expanding their use of BD and AI to analyse demand. Households increasingly adopt technologies that offer tailored recommendations based on their data sharing in pure and mixed data production markets.

Data is not very often traded for money (Economist, 2017) in an explicit fashion. There are several disincentives for trading, such as problems in pricing data, legal restrictions, and the definition of usage in contracts. According to the GDPR (European Commission, 2016), lawful processing ([6]) of data can be based, for example, on consent ([7]) that has to be given for each specific and explicit purpose ([8]). BD knowledge extraction markets are characterised by companies that keep in-house data generation and usage. As a result, when one company is interested in data from another and in its data collection infrastructure, it is likely to buy that company. Bilateral barter deals are also common. The data brokerage industry has also emerged and is characterised by lack of transparency (FTC 2014). It

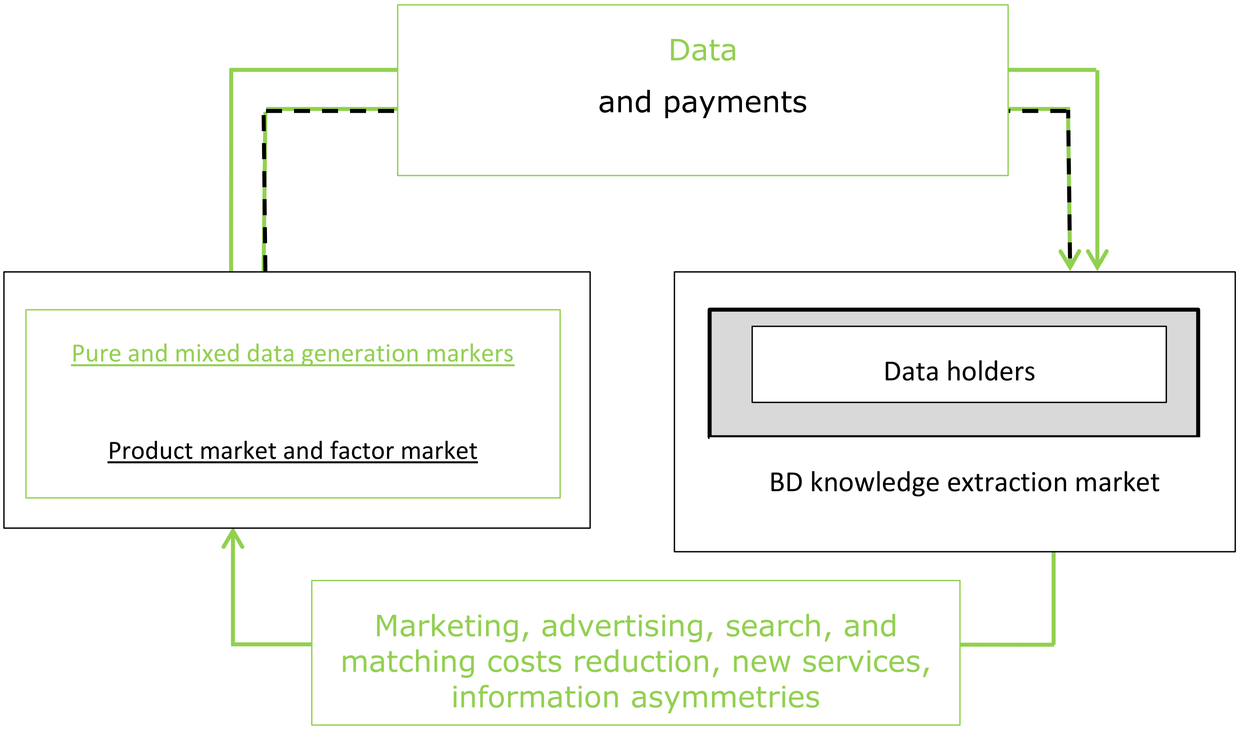
([6]) European Union (2016), Art. 6.
([7]) European Union (2016), Art. 7.
([8]) European Union (2016), Art. 5.b

is composed of unregulated companies that obtain information by scouring web searches, social networks, purchase histories, public records, and other sources (Steel, 2013), very often without consumers' permission. According to Bergemann and Bonatti (2018), firms such as Acxiom, Nielsen, and Oracle sell information about consumers to advertisers and retailers.

The semicircular flow of the economy

Figure 5 takes stock of all of the above to represent the semicircular flow of the economy by incorporating new markets, new flows, and new stakeholders. On the left-hand side, households and firms continue operating in the circular flow of the economy, exchanging goods and services for money and labour for wages. Their daily activity and interactions are increasingly interconnected with pure and mixed data, generating a flow of data and payments to data holders. On the right-hand side, data holders use BD and AI to extract knowledge and information. In return for monetary and data payments, they generate a flow of services that foster aspects of human activity such as sharing of assets, services, and labour in very diverse P2P, B2C, and B2B platforms. As profit maximisation agents, data holders produce the amount of knowledge that maximises their profits. Their services influence markets through matching efficiency, marketing, and advertising, search and transaction costs, creating information friction and generating new and innovative services (OECD, 2019).

Figure 5. The semicircular flow of the economy



It could be argued that data holders are buyers of data inputs and producers of data-driven service outputs, and then fit them into the traditional flow in the box on the left hand side of the diagram. Figure 5 shows that separating the two is important for several reasons. Firstly, the BD knowledge extraction markets are not only a small addition to factor and product markets. "Digitalisation and data flowing is giving shape to the globalisation process and creating a second economy that is vast, automatic, and not easily visible and is not a small add-on to the physical economy. In two to three decades it will surpass the physical economy in size" (Arthur, 2011) and is already able to influence an increasing number of sectors in the economy (Codagnone and Martens 2016; Codagnone, forthcoming; OECD, 2019; Unctad, 2017). Secondly, it visualises the economy as a two-sided market with an intermediary platform. Indeed, the latest papers on this subject point out the implicit new side of markets (Dosis and Sand-Zantman, 2019; Jones and Tonetti, 2018). Thirdly, separation helps visualisation of the semicircular nature of the digital economy, the circular flow assumptions that are failing, and the digital economy market failures that are studied in the following subsection of this paper. Separation based on the role of a factor is also used in this paper to study specific aspects of the circular economy. For example, electricity is an output (of the electricity industry or of households with solar panels) but also an input (into the production of many goods and services). Issues such as concentration in both markets are different and are studied separately. It is known that changes in electricity prices affect costs in the whole economy, which makes concentration in electricity production markets a relevant economic issue. The influence of data in the economy is even more complicated and opaque, which makes the separation relevant and necessary and convenient for human understanding.

Circular flow assumptions that fail in the semicircular flow of the economy

Several assumptions upon which the traditional circular flow of the economy model is built fail in the semicircular flow model.

Money versus data

In the circular flow of the economy, money serves as medium of exchange, unit of account, and store of value. As a means of payment, money is a clear yardstick with a clear value that allows the adjustment mechanism between supply and demand to generate explicit price signals. Money thus facilitates the functioning of the invisible hand: it is easy for everybody to understand and use it and its commonly accepted usage simplifies economic life (Samuelson and Nordhaus, 2010). Central banks control the supply of money and its price in terms of the interest rates.

In pure and mixed data generation markets, each transaction generates information and data but not always a monetary flow. Data take on some of the roles of money. They can be considered to be a **means of payment** (Evans, 2013; Liem and Petropoulos, 2016; Tett, 2018; Scott Morton et al., 2019) but the value of data is neither clear nor set by any authority, and data flows do not generate clearly comparable signals like prices.

According to the *Financial Times* interactive calculator (Steel et al., 2013), data brokers pay between EUR 0.0005 and EUR 0.66 (calculations made in October 2018) for the data of individuals, depending on personal characteristics and the amount of detail. Some consider the data of individuals to be valueless, pointing out that it is only having hundreds of millions of pieces of data to mine and the process of mining it that adds value to the data (Worstall, 2017) by making it possible to predict and influence consumer behaviour. The advertising revenue of data holders illustrates that (Facebook, 2014; Statista, 2015) the value of data is only realised after extracting knowledge from it. Atomistic data is not very valuable but massive data sets are. The positive externalities that data aggregation generates cannot be captured by individual agents, only by large data aggregators. This shifts the marginal productivity of information obtained from data away from individuals towards large computing machines with strong data feeds. Individuals can receive services generated after knowledge has been extracted from the data. Although people are more and more aware of this, many households and firms ignore the amount of data they produce in their daily lives and their (potential) value.

Data cannot play the role of money as a **unit of account**. Several attempts are being made to measure the value of services in pure and mixed data generation markets (Coyle, 2018) and of the data held by data holders (Brynjolfsson et al., 2018; Tett, 2018).

Data, as a key input to innovation (OECD, 2019), can be a potential **store of value** that generates competitive advantages in specific industries. Data can also give inside information about the whole economy and the economic cycle (Artola et al., 2015; Bollen et al., 2011; Choi and Varian, 2011; Askitas and Zimmermann, 2011a, 2011b, 2011c; Gerow and Keane, 2012, Janetzko, 2014), open new opportunities for AI implementations, and generate market power. This makes data generation markets very attractive for investors and venture capitalists (Cadwalladr, 2017).

There is no data authority equivalent to the central banks for the supply of money. An important legal difference between money and data is that whoever owns money can spend it on anything. According to the GDPR (European Commission, 2016), data holders who have received informed consent from a data subject can only use the data for the specific and explicit purpose for which consent has been given.

Circular versus semicircular data flows and no data leakages

Data flows are semicircular: from households and firms to data holders. Household and firms receive data-driven services created from BD & AI but do not receive unprocessed data. In general, unprocessed data is not accessible to consumers, firms, governments, and the public sector. Unlike money, data does not generate a circular flow. Participants in platforms, for example, only receive the information that search algorithms show. The algorithms themselves are a black box. BD knowledge extraction is dominated by data holders that collect BD to feed AI and make more data collection possible. Even if there were a flow of data from data holders to households, they would not generally have the capacity to process them. The role of data and access to and reuse of it are of growing importance (OECD, 2019); data protection regulations and authorities are being set up (European Union, 2016 ([9])) and mainly focus on data protection aspects but not on other dimensions of data. The right to data portability (European Union, 2016 ([10])) has the potential to articulate a circular flow but stills needs further implementation and legal development beyond data protection (De Hert et al., 2018).

Product and factor markets versus the multifaceted role of data

The traditional distinction between product and factor markets is blurred. Data is not only an input but also an **output** of pure and mixed data generation markets. For example, credit card data is an output of consumption processes that data holders collect to understand and predict consumer behaviour more accurately. They are also an **input** in BD knowledge extraction markets used to produce better digital services. Some consider data to be the 'oil of the twenty-first century' (Liem and Petropoulos, 2016) or a critical ingredient in virtually all innovation (OECD, 2019). In fact, as explained above this multifaceted role also applies in the traditional economy for electricity that is an output of the energy production process and an input in other production processes.

Property rights enabling scale and scope

The circular flow of the economy assumes clear property rights for factors of production. However, ownership of data is often unclear (Duch-Brown et al., 2017) and characterised by a de facto situation rather than clear property rights. De facto ownership supports the existence of the black box and market concentration. It constrains alternative uses of data: merit and non-rival uses of data that could increase knowledge and well-being without damaging privacy and the interests of data holders. This makes theoretical categorisation of data fundamental to identifying optimum degrees and forms of data sharing and access (OECD, 2015; Palfrey and Gasser, 2012; Scott Morton et al., 2019). In fact, GDPR (European Commission, 2016) intends to facilitate the free flow of personal data with the goal of protecting the rights of citizens. De Hert et al. (2018) point out the right to data portability is the novel feature of the GDPR that forms the basis for additional regulation beyond data protection and towards competition law or consumer protection.

Specialisation versus sharing

Although many aspects of the economy are still characterised by specialisation, the reduction in entry and search costs makes P2P sharing of idle assets among households possible, for example. Boundaries conventionally drawn between consumption and leisure and production have become more porous (Coyle, 2018; Coyle and Nakamura, 2019): private houses substitute for hotels and private cars substitute for taxis. According to Coyle and Nakamura (2018), digital technologies are bringing about substitution across both market and home production. The lack of specialisation and the switching of economic activity across production boundaries challenges regulations in many aspects: consumer protection, licences, taxation, labour conditions, informal supply of services, quality standards,

([9]) European Union (2016), Art. 51.
([10]) European Union (2016), Art. 20.

professional regulations, unfair competition with formal B2C service providers, and so on (Codagnone and Martens, 2016, Vaughan and Hawksworth, 2014; WEF, 2013; Malhotra and Van Alstyne, 2014) ([11]).

National versus platform jurisdiction

The traditional circular flow framework assumes that activities take place in a country and are taxed and regulated by the rule of law in that country. International activities are regulated by international law, agreements, and institutions in which countries are represented. Data flows in pure and mixed data generation markets operate globally, data holders operate online, and services are more intangible. This facilitates tax avoidance, breaches of the law, and avoiding regulation and state paternalism more easily. There is evidence in the literature of weaknesses in the international corporate tax system (D'Andria, forthcoming). It has been shown that large concentrations of large web-based company profits are found in low-tax jurisdictions whereas the value generated is distributed across countries. Some countries have started to tackle underpayment of taxes unilaterally (Pratley, 2018; Sandle, 2018) and other countries and international institutions have followed and are now devising how to tax digital activity (European Commission, 2018b).

In traditional marketplaces, matching efficiency is driven by supply and demand search efforts, matching technology, information friction, and geography, but there is no specific centralised agent behind it. In market platforms, matching is driven by data holders who take a more active and prominent role that goes beyond supply and demand search efforts and matching. They act as intermediaries in an increasing number of markets by collecting data and obtaining information and knowledge about both sides of the market (Fradkin, 2015). They can influence the search and the matching efficiency, the number of participants who are brought together in the marketplace, and the distribution of information among participants. For example, in P2P sharing, platforms can decide about search rankings, the amount of information made available, exchange conditions, fees, matching algorithms, and regulations that rule the platform's economic activity regardless of where the participants are located. If not too small, as may happen in market platforms with network effects, data holders may be able to influence the behaviour of market participants and violate the natural meaning of competition (Rothschild and Stiglitz, 1976) at their convenience. This may raise concerns about dominance, monopsony-type market power, and abuse of economic dependence in a way that is not easily captured by competition legislation (Codagnone and Martens, 2016). The ability of data holders to collect fees strengthens their financial power ($I\uparrow$) and is accompanied by a reduction in the government's ability to collect taxes ($T\downarrow$) and regulate markets[12].

In summary, economic activity is increasingly characterised by data generation in factor, product, pure, and mixed data generation markets. Table 1 summarizes the circular economy assumptions that fail in the data economy. Payments are not always made using money; data play a multifaceted role as a means of payment, as an input and as an output; neither data ownership nor their value are clear; goods and services are increasingly supplied by unspecialised sellers that depend on data holders; the public sector does not always regulate or establish the rules of marketplaces and online platforms; and digital markets are frequently not located in a specific country and subject to its regulations and property rights. Data flows in the digital economy are not circular and there is no data authority with a role that is equivalent to the role of a central bank in regulating flows of money in the traditional economy. Whether the semicircular flow model presents an economic problem and how regulators may react requires further analyses of its inherent market failures. In the following sections, we explore market failures in the knowledge extraction market black box and their consequences in factor, product, pure, and mixed data generation markets.

([11]) Problems also include measuring gross domestic product, productivity, and welfare, which needs reliable research to inform welfare policy (Coyle, 2018). This is also opening new opportunities that generate new positive externalities (Andrade et al., 2014; WEF, 2014).

[12] And even by some data holders copying structures and institutions of states such as Courts (Macione 2019)

Table 1. Circular flow of the economy assumptions that fail in the data economy

Circular flow model	Semicircular flow model
Monetary flows	Data flows
Circular flow	Semicircular data flow
Products and factors	Multifaceted role of data
Property rights	Unclear data property rights
Specialisation	Sharing
National and international jurisdiction and taxation	Data holders jurisdiction and fees

4. Market failures in big data knowledge extraction markets

Data holders would not have an incentive to invest in pure and mixed data generation markets if they could not obtain at least de facto ownership of data and if there were no other market or production process where they could obtain a revenue stream from data. BD knowledge extraction markets were referred to in the previous section of this paper as a black box. This section examines the characteristics of the process of extracting knowledge information, the existence of economies of scale and scope, sunk costs, network effects, and the economic characterisation of data. Market failures are diverse and specific to each situation. The authors of this paper create explanations using general economic principles and illustrate them with specific examples from daily life, for which there is a need for more empirical evidence in most cases. If knowledge extraction has the characteristics of a natural monopoly, the knowledge generated would be less than socially desirable. Economic characterisation of data supports better regulation of de facto ownership of data, regulation that should maintain incentives for data holders to invest in data markets while at the same time ensuring additional knowledge generation.

Economies of scope and scale in big data knowledge extraction markets

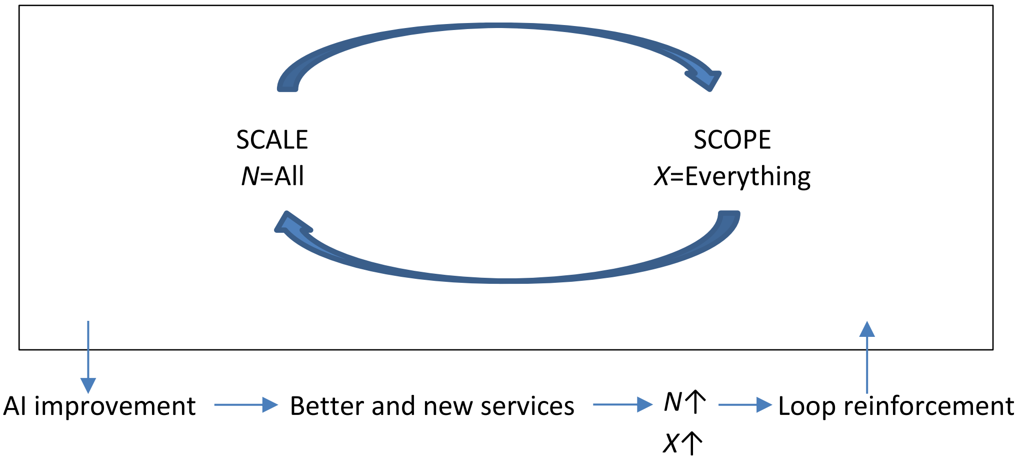
The value of data is released after knowledge or valuable information has been extracted from it. AI offers a scaled-up automated application of existing statistical techniques that enables recognition of patterns, regularities, and structures in data without an a priori theoretical framework (Boisot and Canals, 2004; Duch-Brown et al., 2017; Vigo, 2013). Machine learning models can be tested and continuously improved with new data. Economies of scope and scale arise because data holders have an incentive to centralise their processes of extracting knowledge/information from data and applying it to services such as marketing. Efficiencies are formed by volume (scale) and variety (scope) and involve lowering the average cost of producing knowledge/information from bigger and more detailed data sets in order to improve the prediction of consumer needs and to market more types of products in real time. Scale and scope are a direct consequence of ‘two Vs’ in the definition of BD (Laney, 2001, 2012): volume and variety. Scale and scope explain the emergence and growth of (digital) giant data holders such as Google, Facebook, Apple, Amazon, or eBay and their investment in companies operating in pure and mixed data generation markets, AI research companies, offline products such as mobile devices and gadgets such as smart watches, and services.

Regarding knowledge/information extraction, algorithms trained on one data set may be transposed to other complementary data sets and adjacent data (Duch-Brown et al., 2017) to obtain more and better predictions. The greater the amount of information available about consumers, the better their preferences, and needs can be identified. In statistical terms, scale refers to the number of observations (N) and scope to the number of explanatory variables (X). Volume facilitates the determination of the specification of models because the larger the number of consumers observed (N), the greater the degrees of freedom to include more variables (X). The higher the number of significant variables (scope/variety) and observations (scale/ volume), the more robust and complete are estimates of consumer patterns and behaviour. Scale facilitates scope and scope is an incentive to increase scale. Think of an algorithm to predict where a consumer is planning to spend his or her next holiday in order to offer tailored goods and services. A consumer planning a trip is very likely to search online for the name of the place (Artola et al., 2015). Using online searches as a predictor of holiday destinations, the algorithm can improve by including more variables, such as other searches about hotels, flights, Facebook activity sharing plans with friends, or payment data. The algorithm can also improve by including context variables such as the weather forecast. A posteriori testing is also possible by, for example, using the geographical location of that person in a search engine. In summary, as long as agents are connected and generating data and the use of technology expands, more and more variables can be taken into account to predict behaviour, making the observation of more individuals desirable. Consequently, economies of scale and scope apply to an increasing amount of human activities and sectors of the economy. This is reinforced by the ‘self-teaching’ nature of algorithms when they are fed with more data. This also leads to existing services being improved, new ones developed, and additional data collected on new aspects of life. Looking at categories in the Apple App Store and Google Play illustrates data holder interests in increasing X. They mainly focus on games but expand to lifestyle, education, health, finance, news, and social networking ([13]). Data collection from non-users illustrates interest in increasing N (Financial Times 2019). The acquisition of the Weather Company by IBM in 2015 shows the importance of contextual data.

Similar dynamics apply to knowledge applications to services such as marketing or match-making. Cross-selling one product alongside another, using the outputs of one business as the inputs of another, is an example of economies of scope. Better predictions can improve marketing, reduce search costs, generate synergies, and improve consumer experiences. Economies of scope make product diversification an efficient growth strategy (Ansoff, 1957) if based on common know-how. As similar tools and channels are able to market multiple products more cheaply and efficiently in combination than separately, data holders can reap economies of scope. Offering a range of products tailored to specific consumer characteristics, their situation, personality, needs, contexts, and consumption habits in real time and at the right moment gives consumers a more desirable experience at lower search costs than separate consumption of individual products. Following up on the holiday example, companies can offer car rental or hotel accommodation to consumers who have just booked a flight. Building on this holiday model, data from Spotify, Songkick, Tinder, etc., can match the traveller with events that will interest him or her or with travellers with similar interests.

Figure 6 summarises the scale and scope reinforcing loop that characterises BD knowledge information extraction and its applications to marketing. Data holders are in a race towards *X* = Everything accompanied by *N* = All (Hey et al., 2009; Mayer-Schonberger and Cukier, 2013) enabling self-teaching algorithms, existing services to be improved, and generating new ones that expand the use of BD and AI to new realms of life and increase the number of people from whom data is captured, which reinforces the loop by increasing *N*, people observed, and *X*, aspects of their lives.

Figure 6. Big data knowledge extraction scale and scope loop



Economies of scale operate up to a point at which they may give way to diminishing returns of scale. Thinking in standard statistical terms, this means that the additional knowledge gained by including more observations (*N*) begins to decrease. In a simple ordinary least squares estimation of a dependent variable as a function of several explanatory ones, diminishing returns may begin after a few thousand observations (Varian, 2013). After a certain *N* randomly extracted from the same population, estimated elasticities change very little. That is **not** necessarily the case in knowledge and information extraction from BD in which economies of scale and scope operate together and reinforce each other. As more and more aspects of the lives of consumers can be observed, more predictors can be tested and more services can be 'data driven', making more volume desirable. It is difficult to imagine when diminishing returns may begin in a situation in which almost any variable can be tested as a predictor of the real-time needs of any consumer. In fact, AI algorithms should become more

([13]) See the most popular Apple App Store categories in September 2018 (https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/). See the most popular Google Play categories in September 2018(https://www.appbrain.com/stats/android-market-app-categories).

accurate as more data becomes accessible to them. For example, the more people use Facebook and the more information about them is available in their profiles, the more accurately targeted advertisements should become. The more people use search engines, the better they work. And the more people tag pictures of friends, the higher the accuracy of recognising the faces of people in pictures.

Identifying where economies of scale stop and give way to diminishing returns is an empirical question on which little evidence is available to our knowledge. There is some work showing that the marginal cost of storing an additional megabyte of data is very low (Rubens, 2014). Moreover, AI systems are now being developed that are capable of even generating their own training sets of data without any 'pollution' from established human theory. For example, this was how the entirely self-taught (by playing itself an enormous number of times) AlphaGoZero AI algorithm became the best 'player' of Go in history, far exceeding human capabilities. The hypothesis 'the more data the better' (Silver et al., 2017) without hitting diminishing returns is anecdotally supported by the emergence of a small number of 'massive' data holders in the form of the dominant platform companies (Facebook, Google, Amazon) expanding to new sectors and activities. OECD (2019) and Unctad (2017) report evidence of increasing concentration in the digital sector and lower tax rates. AI- and BD-driven acquisitions and partnerships between companies (Economist, 2017) also illustrate the presence of incentives to centralise BD knowledge extraction and the scale and scope loop.

The costs of diversification and innovation may operate in opposition to scale and scope. They oppose concentration in service, product, and data production markets but not BD knowledge extraction.

Regarding **cost of diversification**, it usually requires a company to acquire new skills, knowledge, resources, and understanding of market behaviour and product development. Extracting knowledge from data to obtain insights into behaviour and marketing, reducing search costs, match-making, advertising, and creating digital services is a specialisation in itself. Data holders are specialised in extracting knowledge and information from BD. They are very often not directly involved in producing physical goods as other agents provide the final goods. There is no initial need to develop new physical products, but they need to innovate continuously to keep data factories (pure and mixed data generation markets) functioning. Platform economy has its limits (Azzellini et al. 2019) but data holder expand into physical production and sectors where platforms are not yet taking over. Such an expansion is mainly data driven and it does not imply a new specialisation. This is the case in Amazon's acquisition of Whole Foods (Hirsch, 2018) or Sofa Sounds partnership with Uber and AirBnB (Azzellini et al. 2019). However, knowledge extraction is not specifically included in the NACE classification ([14]) system.

Regarding **innovation,** it has been traditionally considered as opposing scale and scope. However, the scale and scope loop implies that giant data holders take over new innovations. Successful innovations in pure and mixed data production markets are very often taken over especially when they display networking effects. Instagram and WhatsApp's acquisitions by Facebook in 2012 and 2014 respectively and Google's acquisition of YouTube in 2006 are good examples. Facebook's DeepText, an AI natural language processor able to learn the intentions and context of users in 20 languages and investments made by Facebook in face recognition([15]) technologies show how concentration is accompanied by AI. In general, access to data is a critical ingredient in innovation (OECD, 2019).

Studies in the literature have identified **other forces** that may oppose and limit concentration such as capacity constraints, product differentiation, specialisation, and vertical or horizontal differentiation, congestion, heterogeneity, and multi-homing. There can be capacity constraints regarding advertising space or the variety of products that can be displayed. Technology is increasingly able to overcome some of these problems by improving searching, targeting, and

([14]) The Statistical Classification of Economic Activities in the European Community, commonly referred to as NACE (for the French term *"nomenclature statistique des activités économiques dans la Communauté européenne"*), is the industry standard classification system used in the European Union. It is the European implementation of the UN classification ISIC.

([15]) Mark Zuckerberg's long-term plan for Facebook is centred on three main pillars: artificial intelligence, increased connectivity around the world, and virtual and augmented reality. In November 2016, Facebook acquired FacioMetrics, a face recognition technology company started out of Carnegie Mellon. Face recognition is the new way to identify individual persons, the substitute of fingerprints that will allow real time recognition. It will help in catching criminals but is raising a lot of privacy concerns (Bedoya 2017).

matching algorithms that are able to reduce the problems of capacity constraints. Multi-homing implies that one or both sides of the market use more than one platform. It is limited if there are network and lock-in effects (Evans, 2003). These forces oppose concentration in pure and mixed data generation markets but not the scale and scope loop of BD knowledge extraction. They do not oppose incentives to centralise knowledge extraction and the resulting AI- and BD-driven partnerships and acquisitions. Interconnections and partnerships between agents that operate in different data generation markets illustrate this ([16]).

Sunk costs

Fixed costs in the data economy tend to be especially high. On the one hand, there is connectivity infrastructure for such things as broadband Internet connections that make the Internet accessible to households, firms, and data holders (Unctad, 2017), which are very often developed by the public sector. On the other hand, there is investment in research and development, data centres, cloud computing arms, and data refineries to handle data generation and collection to develop AI and improve knowledge extraction. As AI development is fed with the data, access to data is a key to innovation (UTI, 2018) and a barrier to entry (OECD, 2019). This surely explains at least part of the reason why data holders so often invest in (apparently) non-profitable companies that have developed data generation capacity in pure and mixed data production markets (Bond and Bullock 2019; Kaminska, 2016; McArdle, 2019).

Network effects and barriers to entry

Network effects occur when the number of users increases the value of the service for others using the service. The more people there are with a Facebook profile, the more people can be contacted by Facebook and the better the service provided by Facebook will be. Indirect network effects occur when users on one side of the market attract users from the other side. In a marketplace more buyers attract more sellers, increasing the variety of goods and attracting more buyers. Online platforms are characterised by the existence of direct and indirect network effects and a high fixed costs and low variable costs structure (Duch-Brown, 2017a). This may lead to higher concentration rather than the use of traditional marginal cost pricing.

There might be negative externalities from additional users if this implies an increase in group heterogeneity and in search costs. As a result, vertically and horizontally specialised platforms may emerge with specific target audiences (in, for example, academia, dating, job searching, etc.). These negative externalities may oppose concentration in pure and mixed data generation markets in which different platforms may continue operating separately even if they are technically within the same parent company such as Facebook and WhatsApp. For example, during the process of obtaining European Commission approval to merge Facebook and WhatsApp (European Commission, 2017), Facebook pledged that it would not merge user -bases but has since said that it intends to do so, showing that network negative externalities do not oppose the scale and scope of knowledge extraction.

In general, forces that oppose concentration make markets such as the ‘app economy’ (Basole and Karla, 2012) very specialised and fragmented and generate competition between apps, but, again, this does not necessarily counteract the digital scale and scope loop described in Figure 6 ([17]).

([16]) For example, MasterCard Advisors are IBM Watson partners. PayPal is, in principle, a Mastercard competitor, but Mastercard owns a percentage of PayPal and PayPal is a Facebook partner. Facebook has received investment from PayPal. In China, social networks and the payment industry are already integrated into the same company through the Chinese ‘WeChat’, which, in a single application, offers services such as Instagram, Facebook, and WhatsApp together with payment services. Google’s acquisition of DeepMind, the world AI leader, in 2014 also illustrates the reinforcing nature of BD and AI. DeepMind also has access to public records through its agreement with the United Kingdom’s National Health Service. IBM’s acquisition of the Weather Company in 2015 illustrates that concentration goes beyond personal data to information on variables that determine consumer behaviour.

([17]) For example, the ‘mobile application ecosystem’ comprises an ecosystem orchestrator, mobile application developers, and mobile device owners that are connected through a market platform (Hyrynsalmi et al., 2014). Data holders have created their own application ecosystems, such as Google Play, the Apple App Store and the Microsoft Windows Phone Store, in which they control the transaction infrastructures. Mobile application offers are used as a tool to distinguish one company from its competitors in the mobile devices market (Hyrynsalmi et al., 2012). Therefore, apps markets may operate like competitive markets but within the framework of a data processor-controlled ecosystem. When apps do not display network effects they can be replicated. But whenever an app successfully displays a network effect it tends to become a takeover target for one of the giant data holders. The aforementioned acquisition of WhatsApp by Facebook and YouTube by Google/Alphabet are two examples (Codagnone and Martens, 2016; Rysman, 2009).

If economies of scale and scope, sunk costs, small marginal costs and other barriers to entry operate together, the market may have natural monopoly features. This raises concerns over market concentration (Mckinsey, 2018, 2019) and the need for research on market structure and anti-trust policies (Scott Morton et al., 2019). A small number of data holders have become digital giants that dominate the ability to collect huge amounts of data and extract knowledge information. By 2017 five out of the ten biggest companies in the world were BD- and AI-related companies (Gray, 2017). Their revenue is higher than the gross domestic product of many countries (Investopedia, 2015). Their power extends beyond their revenues because a small number of companies directly or indirectly control the world economy (Vitaly et al., 2011). Studies in the literature have identified that 'superstar effects' (Mckinsey, 2018, 2019; Scott Morton et al., 2019) may be related to access to data (ITU, 2018). Market concentration is not a clearly established empirical fact. It is not clear whether the overall degree of competition has decreased because of digitalisation. On the one hand, large companies and access to deeper data lakes spur investment and innovation in AI methods and avoid duplication of resources. On the other hand, the concentration of research and development inside a small number of very large companies may reduce innovation by creating barriers to entry for smaller, new, and innovative firms. If monopoly theory applies, the amount of knowledge generated maximises the profit made by data holders and would be below a socially desirable amount. Underutilisation of data as a resource would motivate intervention to boost knowledge production: de facto ownership of data protects investment by data holders and innovation by them but limits additional knowledge generation. It is generally assumed that well-defined property rights are a precondition for efficient allocation of resources. In the following sub-section the economic characterisation of data is explored and whether they are a private good whereby the market can lead to an efficient allocation of resources as well as whether data is a merit or a demerit good is also explored.

Data economic characterisation and ownership

Economists classify goods as public, private, club, or common pool goods ([18]) to identify goods that share similar dysfunctions and so may benefit from similar solutions. In general, with private goods ([19]) the market leads to efficient allocation of resources, while for public goods ([20]) government intervention is needed to avoid issues such as free rider ([21]) problems, underproduction, degradation through overuse, and potential destruction ([22]) ([23]). Traditional solutions to these problems are taxation,

---

([18]) Classification emerges from whether a good is excludable or not and whether it is a rival or not. A good is excludable when individuals can be excluded from using it and non-excludable when individuals cannot be excluded. A good is a rival when the amount consumed by an individual reduces the amount available for others and non-rival when the amount consumed by an individual does not reduce the amount available for others. The possible combinations are traditionally presented in a matrix as in Table 2 (Samuelson, 1954).

([19]) A private good is excludable and rival. Examples are food and clothes. The owner of a jacket can exclude other people from using it and two people cannot wear the same jacket at the same time. Producers of private goods can also exclude consumers not willing to pay the price, which makes it possible for them to make a profit. Economic thinking considers that well-defined property rights is a precondition for efficient resource allocation: market forces of supply and demand determine the price and the quantity based on consumer willingness to pay and cost of production to producers. Government intervention as a supplement to the existing legal frameworks defining property rights, how to buy, sell, and enforce contracts, and so on is not necessary.

([20]) Public goods are non-excludable and non-rival, an example of market failure because property rights are not well defined and people do not weigh up all the costs of their actions. Air, the environment, national defence, and street lights are examples. Nobody can be excluded from breathing and the amount a person breathes does not reduce the amount available to others, which generates overutilisation, and pollution.

([21]) The **free rider problem** implies that public goods can be consumed without contribution or payment. Private production cannot reap benefits from their production. Therefore, there are no incentives to provide them through the market. It may lead to underproduction unless they are produced by the public sector. This is the case for national defence from which none can be excluded. Public interventions of this type are represented in circular flows through taxation (leakage) and government expending (injection) on, in this case, national defence.

([22]) **Excessive use** may result in negative externalities such as air pollution or potential destruction of resources. This is known as the tragedy of commons because common pool resources also suffer from this failure. For example, as it is difficult to exclude fishermen from fish stocks, the amount of fish caught by a ship reduces the amount available to others. Fishing grounds allow for a certain amount of fishing after which they may be damaged. In practice, preservation of resources can be enforced by government regulation of access or assigning property rights. Property rights may imply different things such as the right to access, exclude, and sell that reallocate the use of the resources to avoid overutilisation and future destruction. In the case of the environment the extension of property rights has been proven to reduce pollution

([23]) Both the free rider problem and overutilisation can be solved by restricting access and transforming a public good into a club or private goods. For example, technology can encrypt television or radio broadcasting giving access only to members. Club goods are excludable and non-rival. In the satellite television example, non-members can be excluded. The amount of television consumed by one member does not reduce the amount available to other members. Copyrights and patents are legal mechanisms to enforce exclusion for a period of time.

copyrights, patents, regulation of access, and/or assigning property rights. Measures to eliminate market failures are not always satisfactory as they may generate new failures. In this section we explore what kind of good data is in order to shed light on theoretical data ownership consequences.

According to Duch-Brown et al. (2017), under extremely restrictive assumptions, data can be considered a public good: non-excludable and non-rival.

If a public good is one ‘all enjoy in common and each individual’s consumption of such a good leads to no subtractions from any other individual’s consumption of that good’ (Samuelson, 1954, 1955), data is not a public good. Firstly, not ‘all’ agents enjoy access to data collected by data holders ([24]). It can be argued that ‘all’ may enjoy data but only when processed in the form of customised goods or services (Bergemann and Bonatti 2018, Bergemann et al., 2018). Services produced from data, such as search engines and social networks, may fit the above definition. Nobody is, in principle, prevented from using a search engine and how many searches a person consumes does not reduce the amount available to others. However, search engine users are not directly using the data but the services that the search engine owner extracts from them to maximise its profit. Therefore, only certain **fractions of knowledge** extracted from data can be considered to be a public good but not BD as a whole.

Duch-Brown et al.’s (2017) characterisation builds on the **arrow information paradox** assumption: once data is shown to a potential data client/customer, they can no longer be sold because the customer already has the information. However, first, data is not always restricted to the arrow paradox. Data does not need to be shown in its entirety to a data client. Its content can be perfectly understood by means of metadata, small samples, or examples. In practice data is typically **excludable**. Agents are effectively excluded from data collected by data holders.

Table 2. Public, private, club, and common pool goods

	Excludable	**Non-excludable**
Rival	**Private goods/resources** Food, clothes, cars Digital data in BD knowledge extraction market	**Common pool resources** Fish stocks, timber, coal
Non-rival	**Club goods/resources** Traditional examples: cinema, private parks, satellite television Digital data in the data market, BD knowledge extraction	**Public goods/resources** Air, national defence, knowledge, official statistics, lighthouses, street lights Fractions of knowledge and services extracted from data

Regarding, non-rivalry, in the strict sense, the amount of data used by a company does not reduce the amount available to other companies. Data can be used by any number of agents without being depleted, which opens a great many possibilities for generating social gains by data sharing (Jones and Tonetti, 2018). However, giving the competitors of data holders access to data would reduce the profit that can be obtained from data and reduce the incentive to invest and innovate in data production markets and in knowledge extraction. This could reduce the amount of data generated in the economy and the innovation and knowledge it could generate. That could be considered an overutilisation that would lead to a tragedy of commons damaging the data generation process. Current de facto unregulated ownership includes alienation rights and the right to sell, process, and

([24]) Some agents enjoy public access to some data collected by data holders. For example, flight and train schedules are publicly available for everyone to see. Google makes Google Trends available but that is a very small proportion of its data (Artola et al., 2015). Among the academic community, data access is mainly dependent on contacts and the seniority of researchers and data is very often offered subject to non-disclosure agreements. The real situation for social sciences is that at a time when **data dissemination** and **scholar communication** could increase tremendously, the replicability of research is really in danger. Academic journals are yet to implement a policy of not publishing research carried out using data invisible to everyone but the authors. Companies may ‘not like the prospect of negotiating access with individuals in case by case basis, and decide not to make data available to everybody or to nobody’ (Taylor et al., 2014, page 8).

obtain profit and avoids free-riding from competitors. It preserves incentives for data holders but reinforces market concentration, keeps the ‘black box’ closed, generates information asymmetries, and limits additional generation of knowledge. Rivalry or non-rivalry has traditionally depended on the nature of the good rather than in the market setting. In the case of data, it seems to depend on the type of use and whether it ‘rivals’ the interests of data holders and the data generation processes.

If data is excludable and non-rival, it may be best be considered to be a ‘club’ good within the data lake of a given data processor. De facto ownership avoids problems such as free-riding but generates and reinforces market failures such as **market concentration** and data **underutilisation.**

There are goods that although not exactly within the concept of public goods, can be underconsumed if provided by the free market (merit goods such as education) or overconsumed (demerit goods such as illicit drugs) (Musgrave, 1959). The idea behind merit and demerit goods is that a well-informed society is in a better position to identify the amount of certain goods needed than individual agents and their willingness or ability to pay for and/or process information ([25]). Governments impose community standards and support consumption of merit goods such as education, and ban demerit goods such as illicit drugs. The public sector aims to solve **information failures** and improve consumer sovereignty and is based on knowledge of the individual and community consequences of merit and demerit goods. Depending on how it is used, data can be a merit or a demerit good. Data is a merit good when used to reduce market frictions, information costs, and information asymmetries, to generate knowledge and better matches between supply and demand, to facilitate full performance of private assets that otherwise would be idle, to innovate, to deliver nimbler policies to prevent and mitigate the consequences for the economic cycle, and so on. Data is a demerit good when used to generate market power, set barriers to entry, generate information asymmetries, and when used by a monopoly to control marketplaces, to charge unfair fees or prices, or to impose excessive regulations limiting innovation or generating unacceptable distribution of wealth.

A society where BD knowledge extraction occurs within a black box is not a well-informed society regarding the basis up which many important economic decisions are made. There are potential non-rival and merit uses of data that may increase well-being without damaging data holders and their data factories. The next section focuses on information asymmetries to further identify merit and demerit uses of data and illustrate whether the current flow of knowledge generated by data holders is a socially desirable amount.

([25]) For example, without education and maturity individuals cannot make well-informed choices about the amount of education they should consume. Before education, individuals ignore the positive consequences of higher educational levels on income and happiness. Society as a whole benefits from individual education because it displays positive externalities for well-being, citizen security, and economic growth. Similarly, drug addicts cannot decide for themselves and drug markets and consumption generate negative burdens on the whole of society through expenditure on health and social security.

5. Information asymmetries in factor, product, pure, and mixed data generation markets

Distribution of information has always affected market outcomes (Duch-Brown, 2017a; Stiglitz, 2001), but data ownership, access, and trade in the digital economy may be even more important for economic welfare (Duch-Brown, 2017 a, 2017 b , and 2017 c). Although digitalisation reduces information costs, it does not solve information asymmetries. Data holders do not disclose data or important metrics (Codagnone and Martens, 2016; Hall and Krueger, 2015). Information asymmetries occur between data holders and households (consumers), offering opportunities for price discrimination: between data holders and firms (traditional suppliers) giving rise to concerns about unfair competition, and between data holders and merit users of data, such as regulatory authorities and the social sciences community. In addition, information asymmetries are giving rise to concerns related to inequality in the wider economy and society.

Asymmetric information between data holders and households (consumers)

This asymmetry can generate at least two types of market failures: (price) discrimination and steered consumption. According to a 2015 report from the White House Council of Economic Advisers, 'sellers are now using big data and digital technology to explore consumer demand, to steer consumers towards particular products, to create targeted advertising and marketing offers, and in a more limited and experimental fashion, to set personalized prices' (White House, 2015, Ursu, 2015). Information asymmetries between supply and demand have traditionally occurred in the uncertainty of product quality (Akerlof, 1970). New information asymmetries are arising as a result of the ability of data holders to collect and process consumer data. It creates an imbalance of power in transactions because one side of the market is able to price discriminate by obtaining person-specific reservation prices. There are several pieces of empirical evidence that claim openness and transparency in this respect. Mikians et al. (2012) demonstrated the existence of both price and search discrimination on the Internet. Shiller (2014) found evidence in the Netflix context ([26]). Chen et al. (2015) studied Uber ([27]) price surges, and Möhlmann and Zalmanson (2017) investigated how Uber's drivers may generate price increases, while Uber (2018) claims to be able to self-regulate such situations. Ezrachi and Stuke (2016) showed that sellers are able to find the right emotional moment to steer consumers into buying.

Price discrimination may also be beneficial for more price-sensitive consumers to whom companies can offer the cheapest option. Choey et al. (2016) showed that firms are not necessarily better off when price discriminating than when competing under uniform pricing. Discriminatory pricing may generate competitive reactions opposing it. Consumers have access to more information, search engines, and price robots to find the best offers (Kramer and Kalka, 2016). However, it is difficult for consumers to understand how these search robots work, their reliability, or the algorithms and reasons behind low monetary prices. In some sectors, undercutting monetary prices is the strategy used by new business models to eliminate competition in the supply of traditional products while receiving part of the payment in data.

Tools for discrimination can go beyond prices and lead to discrimination and unfair treatment. Uber, for example, has developed an internal tool called Greyball which uses data collected from the Uber app, geolocation data, credit card information, social media accounts, and other data points to avoid giving rides to certain individuals such as law enforcement officers and government officials in areas where its service is or was illegal (Isaac, 2017; Wong, 2017). Discrimination may be even more problematic in the case of insurance services, which can become prohibitively expensive for people with specific diseases such as genetic disorders.

The potential to discriminate, steer consumers, and identify their sentiments (Pilaszy and Tikk, 2009) at specific moments is an additional incentive for developing scale and scope and for data-

([26]) According to Shiller et al. (2012), using demographics to tailor prices raises profits by 0.8 %. Including nearly 5 000 website browsing explanatory variables increases profits by 12.2 % as a result of some consumers paying double the price others do for exactly the same product.

([27]) Uber Technologies Inc. is a transport network company operating in many cities worldwide through the Uber car transport and food delivery mobile apps. Uber has been a pioneer in the sharing economy. 'Uberification' or 'Uberisation' refers to changes in industries that challenge traditional specialisation as a result of the sharing economy. Uber has been the subject of protests and legal action on several issues because of its, perhaps unfair, competition with some traditional public transport services.

driven partnerships. For example, think about the possibilities of merging data from Facebook ([28]), which has already raised concerns about privacy and micro-targeting of adverts in political campaigns, and payment industry information ([29]), which includes detailed information about spending habits. More knowledge in this respect is socially desirable.

Asymmetric information and unfair competition between data holders and firms (traditional suppliers)

This asymmetry can generate at least two types of market failures: predatory pricing and monopsony behaviours.

Unfair competition emerges as a result of data access and competition via price undercutting leading to predatory (monetary) pricing, while part of the payment is implicitly made using the consumers' data. Predatory pricing implies selling at prices below the cost of production to drive competitors out of the market or to create barriers to entry for potential new competitors so that the predator company becomes a monopoly. Economic theory distinguishes two stages of predation. First, there is the stage when the predator offers goods and services below their cost of production. The predator needs to be financially strong because during this stage the company may incur losses. The second is the recouping stage, which begins once the predator has market power, and the ability to raise prices above competitive, or even monopoly, levels. During this stage it recovers from losses incurred during the predation stage. Predatory pricing may fail if the predator's competitors are strong enough to survive or are replaced by others. The strategy succeeds when the predator is stronger than its competitors and when there are barriers that prevent new entrants joining the market. Empirically, it may be difficult to identify when prices are low because of deliberate predatory pricing rather than as a result of legitimate competition from a more efficient and innovative producer (Bensinger, 2012; Bond and Bullock, 2019; Kaminska, 2016; McArdle, 2019,). In some cases, such as multi-sided platforms, pricing below marginal cost on one side may not be predatory but profit maximisation (Codagnone and Martens, 2016). Predator identification is even more difficult in pure and mixed data production markets when data is part of the payment. Data gives digital predators a stronger position and more chances of succeeding. The predator can move simultaneously to the recouping stage as value of the data can be released in the BD knowledge extraction markets ([30]). More knowledge aiming to improve understanding of pricing strategies in the black box is socially desirable.

([28]) Facebook makes it possible for marketers to effectively target very specific audiences depending on the marketing objectives of advertisements. Facebook's 'topic data' gives marketing personnel information about topics which people are engaged in and enable marketing based on what audiences are saying on Facebook about events, brands, subjects, activities, their sentiments, the volume, the location, etc. This helps create content and identify the perfect time and location to reach potential consumers. Although querying 'topic data' results cannot instantly turn into advertisement targeting, advertisements can be set to target people in similar demographics and with similar variables. To develop 'topic data' Facebook worked with DataSift and its partner NetBase, a social analytics platform, which connects global companies with consumers. NetBase's platform processes media posts for business insights, marketing, customer services, sales, and product innovation. Companies such as American Airlines, Arby's, Coca-Cola, Ogilvy, T-Mobile, Universal Music Group, Walmart, and YUM! Brands are NetBase partners. Advertising on Facebook and topic data have also been used for political marketing. Both tools were used during the Brexit referendum and the 2017 United States electoral campaigns (Cadwalladr, 2017).

([29]) For example, Mastercard holds billions of transactional and payment data on the underlying processes (geographical location, time, online/offline, amount, etc.), the demographic characteristics and the spending habits of cardholders. MasterCard Advisors translates data into behavioural insights and customises services to financial institutions, merchants, media companies, and governments to help them market their products. It provides critical information to direct the right messages and offers to the cardholders most likely to respond.

([30]) De facto ownership of data not only builds up a barrier to entry for new competitors but also for services traditionally provided by the public sector. For example, Kutsuplus was a public version of Uber, developed by Helsinki County Council. Despite year on year growth of 60 %, it shut down because it was too expensive for the local authority to run. The Kutsuplus example raises doubts about Uber's price undercutting policies. Uber is a not profitable company but it is a good example of financial strength. It has the strategic support of data holders and data-hungry venture capitalists such as Google, Amazon, and Goldman Sachs. In the case of public transport in Helsinki, public provision of an Uber-like platform was not viable because of strategic price undercutting. financed by Uber's data holders. This example also illustrates how the economy evolves towards disequilibrium, as in equation 4 in Section 6. First, the public sector shrinks, given the difficulties in taxing digital activity. Second, the digital supplier has enough financial strength to operate with predatory pricing while collecting more data that will increase its competitive advantage and capacity to innovate. In the new equilibrium, public expenditure on public transport is substituted by private investment. This example also illustrates economies of scale and scope. Scale and scope are not limited to an industry or sector but to many realms in the economy and perhaps even the whole economy: N = All and X = Everything. It shows how search engines and online platforms can expand to the transport industry. The expansion goes beyond that, as Uber is evolving from a platform providing a transport network for people to one involved in adjacent industries such as parcel and food delivery, and even towards self-driving cars.

Asymmetric information and unfair competition between **data holders** and **firms** may also evolve by means of **monopsony behaviours**. As the Internet penetrates more activities in society, supply increasingly depends on digital channels to market and sell products. Market structures have emerged in which either a buyer substantially controls the market as the major purchaser or a platform dominates a market. This is the case for Netflix in the film industry or Amazon's dominant position in online book sales (Krugman, 2014). As stated above, the Amazon example illustrates that scale and scope do not refer to a single industry but to N = All and X = Everything. It recently acquired Whole Foods, expanding into fresh food delivery and acquiring a lot of data on the Whole Food shopping experience and obtaining insights into offline/online habits (Hirsch, 2018). In 2016, 30 % of Amazon's profit was generated by its activity as an online retailer. In 2018, that proportion grew to 50 %, which shows its expansion as a marketplace. More knowledge aiming at a better understanding of the market structure of knowledge extraction is socially desirable.

Asymmetric information between data holders and merit users

Information failures and asymmetries in many markets have been substantially reduced. For example, families have more information about other families and the assets they are willing to share, making new exchanges and transactions possible. In general, data holders reduce information and search costs which in principle lead to an improvement in supply- and demand-matching technology (Coles and Smith, 1996) and benefits for all: consumers get more convenient and cheaper choices and producers obtain revenue (Martens, 2016). However, it is not known whether platforms are designed to minimise search costs and maximise user interests rather than to maximise their profit and take advantage of market failures. The flow of knowledge and information in the semicircular flow of the economy mainly relies on the good intentions (Einav and Levin 2013) of the data holders rather than on legal security, law enforcement, and policy action guided by scientific evidence. Agents such as **regulatory authorities** and the **scientific community** could in theory make non-rival and merit use of BD and AI by looking into 'the black box' to guide policy. However, their access to algorithms and data is very limited (Scott Morton et al., 2019; Taylor et al., 2014; Butler 2013; Artola et al 2015; Lazer et al.2014). From a regulatory point of view, the current situation is characterised by many legal voids and the absence of applicable regulations. It sustains itself in an economic sense because data holders have generated their own systems to substitute state regulations, law enforcement, and taxation. Platforms can set transaction rules and conditions and charge fees for their matching services. The 'sharing economy', for example, has developed its own liability systems by means of consumer reviews or ratings. Trust among participants acts as a substitute for consumer protection regulations. These liability systems may reinforce unfair competition because reputation is platform specific, and it is difficult to transfer to a different platform and generate lock-in effects. More knowledge is socially desirable to guarantee data holders act with good intentions.

The **scientific community** has shown the ability of BD to predict a diverse range of real-life variables related to the economic cycle and the stock exchange (Askitas and Zimmermann, 2011a, 2011b, 2011c; Bollen et al., 2011; Choi and Varian, 2011; Gerow and Keane, 2012; Edelman 2012, Pedraza et al., 2019; Reips and Garaizar, 2011) even when the amount of data available to researchers is only a small proportion of data holders' data lakes (Artola et al., 2015, Butler, 2013; Lazer et al.2014). From a scientific point of view, BD are also **underutilised** because society and policymaking are not fully benefiting from the extraction of scientific knowledge that could inform government responses to the economic cycle **macroeconomic stability**. Such use would be non-rival, Pareto efficient, and merit, and could change the scale and scope of knowledge about many phenomena (Schroeder and Cowls, 2014).

Asymmetric information, income distribution, and separation of powers

Distribution of information generates asymmetries, which foster inequalities (Duch-Brown, 2017; Stiglitz, 2001). The unequal distribution of BD knowledge extraction exacerbates inequality. The United Nations (UN, 2013) has called for a global partnership to eradicate poverty and a data revolution to improve the quality of statistics available to citizens and governments. The United Nations (UN, 2014) and the International Telecommunications Union (UTI, 2018) also reported the huge and growing inequalities in access to data, information, and the ability to use it in a world where data is 'the lifeblood of decision-making and the raw material for accountability'. Although focusing on poverty and development and not on the whole second, or digital, economy, the United Nations (UN, 2014) set out the main opportunities and risks presented by the data revolution and warned that information gaps between the private and public sectors can widen abuses of human rights.

Investors with partnerships and the ability to finance data holders may have **inside information** about the whole economy. This may be the case for hedge funds operating in markets around the world and employing AI models fed with as much data as possible. These companies treasure BD, AI, and human intelligence. They recruit and retain very talented scientific staff who have to sign very stringent iron-clad non-competition and non-disclosure agreements. BD and AI know-how has recently been used in electoral campaigns. The intensive use of paid social media marketing may have influenced several political processes in a decisive way (Grassegger. and Krogerus 2017; Kosinski et al, 2013; Cadwalladr, 2017), with voter targeting decisions at least partly based on BD and AI insights. The same agents behind the electoral democratic process and the stock exchange may generate a situation that resembles the separation of powers problem (Kee 2018). The situation seems even more worrying if it is occurring within a black box. This lack of separation may increase the redistribution of wealth towards data holders.

6. Policy option for merit, Pareto efficient and non-rival uses of data: a data authority

In the context of digitalisation, how can the public sector continue playing its role of **responding to market failures and inefficiencies and promoting macroeconomic stability, growth, and equity?**

According to the traditional approach, and not including the role of data and focusing only on traditional flows in the economy, digitalisation decreases the ability of governments to collect taxes ($T\downarrow$) (D'Andria, forthcoming). As a result, the economy will tend to regain equilibrium at the expense of government spending ($G\downarrow$). At the same time, the ability of data holders to collect fees and monetise the value of data and their financial power facilitate their investment ($I\uparrow$) and their expansion into other sectors, even those traditionally publicly funded such as health, education, and public transport. This implies that the role of the state in the economy is, in general, lower.

$$S + T\downarrow + M = I\uparrow + G\downarrow + X \qquad (4)$$

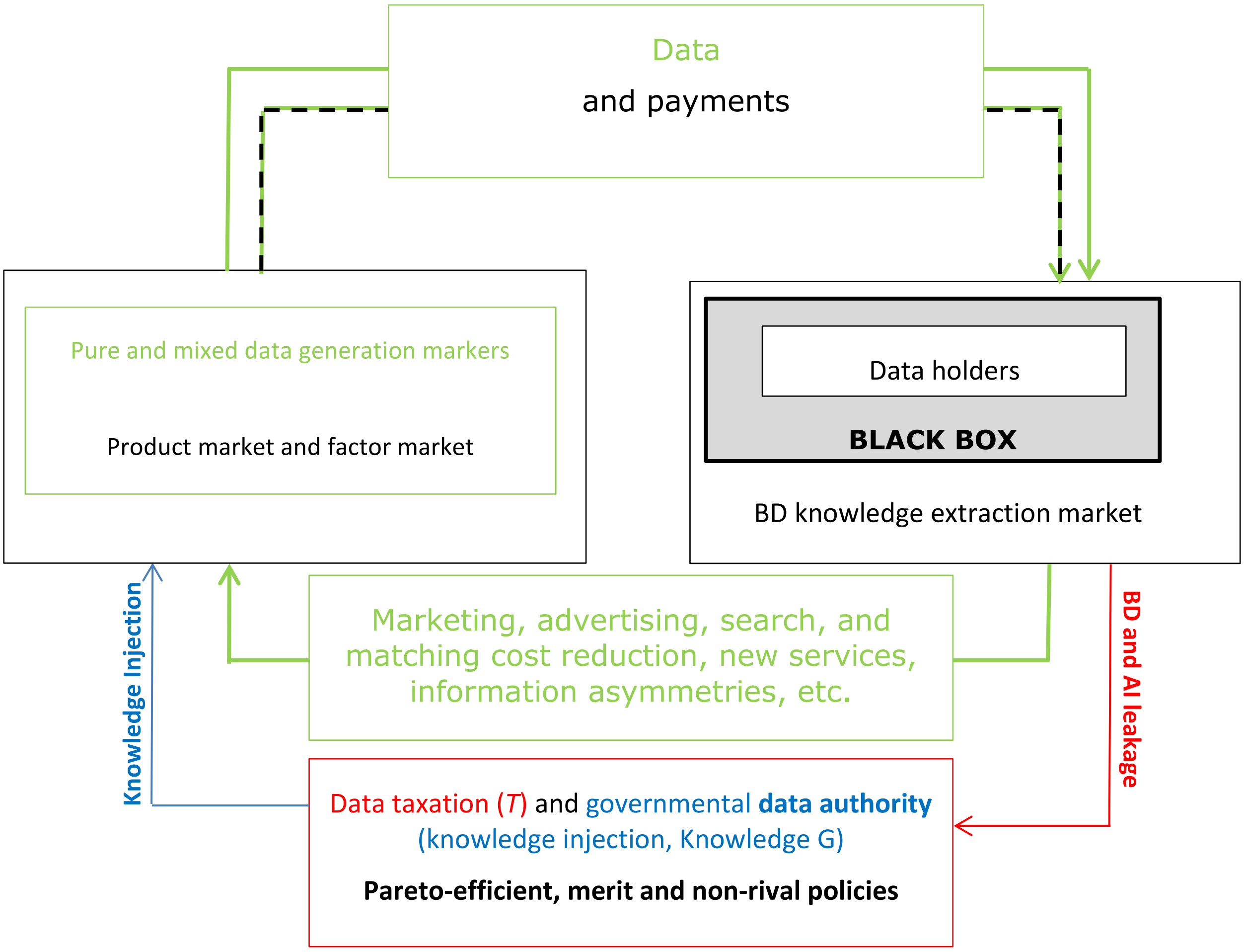
Existing policy actions build on this view of the economy and focus on fostering monetary taxation of digital activities (D'Andria, forthcoming; European Commission, 2018b; Pratley, 2018; Sandle, 2018; Khan and Brunsden, 2018; Munoz de Bustillo, 2019) or fines (Onfro and Browne 2018) to balance the equilibrium in equation 4 without the need to reduce government expenditure ($G\downarrow$). This approach, although probably necessary to guarantee that the state continues playing a role in the digital economy, leaves the public sector displaced and **outside the black box and BD knowledge extraction.** It does not solve data underutilisation or information asymmetries, and does not offer any potential for merit uses of data. The state continues to play its role in the economy without taking into account data and knowledge flows.

Several alternatives have started to appear in the literature. Posner and Weyl (2018) propose that agents could be compensated by the data they generate just as they are compensated for their labour or in the form of a dividend (Ulloa, 2019). Such compensation still has a 'monetary' view of the economy and does not take into account difficulties in pricing data and therefore the amount of compensation. Alternatively, Jones and Tonetti (2019) propose giving data property rights to consumers. However, individual data is almost valueless (Steel et al., 2013) as only having hundreds of millions adds value to data (Worstall, 2017), and the value of data is only realised after extracting knowledge from it. Jones and Tonetti's (2019) proposal could be accompanied by tools and infrastructure that enable citizens to benefit from the positive externalities of data aggregation and knowledge extraction. This leads to the idea of the data authority, as proposed by Martens (2016) and more recently by Scott Morton et al. (2019). Incorporating leakages and injections into the semicircular flow of the economy leads to a similar proposal (Figure 7). A leakage in a data-intensive economy would be a flow of data to a data authority able to generate injections of knowledge on market failures, consumer rights, market structures, inefficiencies, macroeconomic stability, growth, and equity without damaging privacy. Public intervention in the semicircular flow of the economy should aim to achieve the optimal degree of knowledge extraction from data and avoid under-consumption of data as a merit good while reducing, regulating, or banning demerit uses.

Therefore, fostering merit uses of data could be achieved by two approaches: one based on the state's monopoly of power, the other based on data ownership by individuals. The first approach could be executed through data taxation, which would generate a data flow to a data authority that would coordinate knowledge generation. Data taxation (data T) would be the leakage from data holders to the government sector responsible for fulfilling the goal of the governments. The injection would be the welfare knowledge transfers to the community (knowledge G) to promote efficiency, equality, stability, and law enforcement (Figure 7). The second approach would imply that the data flow would go first to individuals, acknowledging that the personal data of an individual is their own property. This option would, in principle, generate less knowledge because data aggregation generates positive externalities that only large data aggregators with large computing machines can capture, taking marginal productivity away from individuals. In order to resolve this, individuals could decide whether or not to share data with the data authority. On the one hand, companies in general probably prefer to negotiate access with only one agent, an authority, rather than with individuals on a case-by-case basis. On the other hand, treating data as the property of an individual would be more socially acceptable and less likely to prompt privacy, surveillance, and 'big brother'-type concerns.

Both approaches are in line with the GDPR's (European Commission, 2016) extensive view of data portability ([31]) (De Hert et al., 2018) according to which data can be transferred from one controller to another. In the second approach, transfers to a data authority could be based on the consent of an individual ([32]). In the first approach, it would be the role of the corresponding public authority ([33]), which would go beyond data protection and towards competition law and consumer protection (De Hert et al., 2018).

Figure 7. Semicircular flow of the economy, leakages, and injections



In any event, data policy should be accompanied by a coordinated **infrastructure,** a centralised scientific authority big enough to benefit from scale and scope in knowledge extraction and benefit individuals through positive externalities of aggregation. Policy reactions have so far not been coordinated, centralised, or carefully designed. Access to data and algorithms are limited and mainly based on individual, case specific agreements (Pawelke and Tatevossian, 2013; Einav and Levin 2013, Barzic et al. 2018, Connolly, 2016; Scott and Young, 2018; Taylor et al., 2014; Prewitt, 2013).

Any government intervention has to be **merit, Pareto efficient, and non-rival.** Generating knowledge flow should foster better regulations, standards, transparency, and maximise the common good (European Commission, 2017), working out information asymmetries and natural monopoly dynamics, aiming to achieve a well-informed society but avoid privacy and data protection issues and

([31]) European Union (2016), Art. 20.
([32]) European Union (2016), Art. 7.
([33]) European Union (2016), Art 51.

free-riding behaviour by the competitors of data holders. Merit and non-rival agents are organisations such as central banks, anti-trust and consumer rights authorities, the scientific community, and other agents who are not are not data holder competitors. Central banks could improve their forecasting of the economic cycle to deliver faster and nimbler policies. Anti-trust and consumer rights authorities could trigger research on sources of market failures, deliver better anti-trust policies, and balance information asymmetries. The scientific community could change the scale and scope of knowledge about a great many research topics and phenomena (Schroeder and Cowls, 2014). New scientific evidence, conceptualisations, and theories would develop theoretical-empirical synergies that would disentangle the reality behind the black box**. Merit knowledge generation from data** would have spillover effects on the whole of society, including data holders. Either data taxation or treating data as the property of individuals with an infrastructure would be Pareto efficient because it would improve the situation of agents who are the beneficiaries of the intervention, mainly households and firms, without generating negative consequences for efficient allocation of resources. It would also avoid problems emerging that are already solved by the current de facto ownership (e.g., incentives to innovate, free rider problems, and the tragedy of commons).

A very sensitive issue in such a data policy would be **privacy**. Any data policy should be implemented in accordance with existing regulation such as the GDPR. The data authority would be in charge of coordinating merit access to data and avoiding privacy issues. Facilitating data access under non-disclosure agreements to merit and non-rival agents is just an additional tool to grant privacy and other citizens and consumer rights and to fulfil the requirements of the GDPR. This is in line with the extensive user-centric view of data portability (De Hert et al., 2018). The explanation above assumes that technology and careful regulation can protect privacy and allow additional generation of knowledge without raising concerns about privacy.

There are a number of decisions and regulations that a data policy would need. To name but a few: revisit the NACE classification to include knowledge extraction from BD as a specific activity, develop professional deontological codes for merit users, and run communication campaigns and provide education to inform the public about the differences between the role of a data authority and activities that endanger privacy such as surveillance, ‘spying on citizens’, and ‘big brother’-type proposals. The data authority would have to be clearly set apart from former scandals such as the disclosure of the US National Security Agency’s mass surveillance and the public outcry following data releases.

7. Conclusions

The semicircular flow of the economy represents the major exchanges of the digital economy. Data flows from firms and households towards data holders. Flows of processed knowledge and information go back to economic agents in the form of 'algorithmic' services. The activity of extracting knowledge from data by means of AI displays natural monopoly characteristics that cast doubt on whether the quantity of knowledge generated is below the perfect competition/socially optimal amount. Knowledge extraction occurs within a black box that produces the amount of knowledge that maximises data holders' profits, causing information asymmetries and inequalities in access to data. There is a lack of transparency and empirical evidences on how the digital economy works. Merit uses of data could activate knowledge flows and shed light on whether government should play a role. Existing regulation in Europe is intended to facilitate the free flow of data within the EU to protect the rights of citizens. The semicircular flow model supports the development of user-centric regulations, data portability, data taxation, and a data authority.

Traditionally, if a monopoly led to higher prices and lower output (lower generation of knowledge), governments would intervene, and if it implied an unfair distribution of wealth (and knowledge and information), the state would implement redistribution policies. Nowadays fiscal and monetary policies are unable to disentangle what happens inside the black box and the consequences for the economy. The semicircular flow defines new flows from which new leakages and knowledge injections can be defined. It motivates the development of a data policy able to work in that direction. It sheds light on how additional regulation could build on GDPR's portability and free flow of data.

Breaking up monopolies may generate inefficiencies arising from duplication of resources. The economic characterisation of data identifies dysfunctions and possible solutions and where intervention could display positive externalities without reducing efficiency. Data characterisation supports further clarity in the de facto ownership of data holders by distinguishing between data use (by data holders and merit users) and data property rights (of consumers). On the one hand, clear rights should maintain data holders' incentives to invest and the efficiencies that emerge from positive externalities of aggregation and avoid free riding, overutilisation, and the tragedy of commons that would damage data generation markets. On the other hand, allowing additional merit uses of data should increase the amount of knowledge generated, transparency, and market efficiency. Exploring the information asymmetries between data holders and the rest of the agents in the economy helps to identify underutilisation and uses that would allow policy measures to eliminate market failures without generating new ones.

Data policy should also aim to increase the amount of knowledge generated by studying data and algorithms in the 'black box' to enable governments to play their role in the digital economy**: responding to market failures, macroeconomic stability, inefficient growth, and equity**. Merit uses would support consumer and anti-trust authorities, central banks, data protection agencies, and the scientific community. Arguments, such as scale and scope and access to data, support a centralised data authority with enough infrastructure and resources.

The main contribution of the semicircular flow of the economy is to facilitate a simple economic theoretical motivation for access to and portability of data for the public good. Many questions for future research emerge. They range from new theories and empirical evidence to policy actions. A data policy should foster interaction between science and policy to reinforce each other. Future research by the authors will focus on identifying the optimum degree of data utilisation and optimum data taxation and the micro foundations of the semicircular flow of the economy.

References

Akerlof, G. A., 1970, 'The market for "lemons": quality uncertainty and the market mechanism', *Quarterly Journal of Economics*, Vol. 84, No 3, pp. 488-500.

Anderson, C., 2008, 'The end of theory: the data deluge makes the scientific method obsolete' *Wired*, 23 June 2008 (http://www.wired.com/science/discoveries/magazine/16-07/pb_theory) accessed 24 September 2018.

Andrade, P. L., Hemerly, J., Recalde, G. and Ryan, P., 2014, 'From big data to big social and economic opportunities: which policies will lead to leveraging data-driven innovation's potential?', in Bilbao-Osorio, B., Dutta, S. and Lanvin, B. (eds) *The global information technology report 2014*, World Economic Forum and INSEAD, Geneva.

Ansoff, I., 1957, 'Strategies for diversification', *Harvard Business Review*, Vol. 35 No 5, pp. 113-124.

Arthur, W. B., 2011, 'The second economy' *Mackinsey Quarterly*, October 2011 (http://www.mckinsey.com/insights/strategy/the_second_economy).

Artola, C. and Galan, E., 2012, *Tracking the future of the web: constructing of leading indicators using internet searches*, Banco de España, Documentos Ocasionales No 1203 (http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf).

Artola, C., Pinto, F. and de Pedraza, P., 2015, 'Can internet searches forecast tourism inflows?', *International Journal of Manpower*, Vol. 36, No 1, pp. 103-116 (http://www.emeraldinsight.com/doi/pdfplus/10.1108/IJM-12-2014-0259).

Askitas, N. and Zimmermann, K. F., 2009, *Google econometrics and unemployment forecasting*, IZA Discussion Paper No 4201, June 2009, Institute of Labor Economics.

Askitas, N. and Zimmermann, K. F., 2011a, *Health and well-being in the crisis*, IZA Discussion Paper No 5601, March 2011, Institute of Labor Economics.

Askitas, N. and Zimmermann, K. F., 2011b, *Detecting mortgage delinquencies*, IZA Discussion Paper No 5895, July 2011, Institute of Labor Economics.

Askitas, N. and Zimmermann, K. F., 2011c, *Nowcasting business cycles using toll data*, IZA Discussion Paper No 5522, February 2011, Institute of Labor Economics.

Azzellini, D., Greer, I., Umney, C. 2019. *Limits of the Platform Economy: Digitalization and Marketization in Live Music*. Working Paper Forschungsforderung number 154, August 2019, Hans Blockler Stiftung.

Barzic, G., Rose, M. and Rosemain, M., 2018, 'French officials are going to work at Facebook for 6 months' World Economic Forum (https://www.weforum.org/agenda/2018/11/france-to-embed-regulators-at-facebook-to-combat-hate-speech/).

Basole, R. C. and Karla, J., 2012, 'Value transformation in the mobile service ecosystem: a study of app store emergence and growth', *Service Science*, Vol. 4, No 1, pp. 24-41 (https://doi.org/10.1287/serv.1120.0004).

Bensinger, G., 2012, 'In Kozmo.com's failure, lessons for same-day delivery', *Wall Street Journal*, 2 December (https://blogs.wsj.com/digits/2012/12/03/in-kozmo-coms-failure-lessons-for-same-day-delivery).

Bergemann, D. and Bonatti, A., 2018, *Market for information: an introduction*, Cowles Foundation Discussion paper No 2142 (http://www.mit.edu/~bonatti/infointro.pdf).

Bergemann, D., Bonatti, A. and Smolin, A., 2018, 'The design and price information', *American Economic Review*, Vol. 108, No 1, pp. 1-48 (https://www.aeaweb.org/articles?id=10.1257/aer.20161079).

Blake, T., Nosko, C. and Tadelis, S., 2014, *Consumer heterogeneity and paid search effectiveness: a large scale field experiment*, NBER Working Paper 20171 (https://www.nber.org/papers/w20171.pdf).

Boisot, M. and Canals, A., 2004, 'Data, information and knowledge: have we got it right?', *Journal of Evolutionary Economics*, Vol. 14, No 1, pp. 43-67 (DOI: 10.1007/s00191-003-0181-9).

Bollen, J., Maoa, H. and Zengb, X., 2011, Twitter mood predicts the stock market. Preprint submitted to *Journal of Computational Science*, 2(1), March 2011, pages 1-8.

Bond, S. and Bullock, N., 2019, 'Uber IPO prospectus shows ride-hailing revenues stalled', *Financial Times*, 11 April (https://www.ft.com/content/c68d3662-5c76-11e9-939a-341f5ada9d40).

Brynjolfsson, E., Eggers, F. and Gannamaneni, A., 2018, *Using massive online choice experiments to measure changes in well-being*, NBER Working Paper 24514 (http://www.nber.org/papers/w24514).

Butler, D., 2013, 'When Google got flu wrong', *Nature*, Vol. 494, 14 February 2013.

Cadwalladr, C., 2017, 'Robert Mercer: the big data billionaire waging war on mainstream media', *The Guardian*, 26 February 2017 (https://www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage) accessed April 2017.

Chen, L., Mislove, A. and Wilson, C., 2015, *Peeking beneath the hood of Uber*, (DOI: http://dx.doi.org/10.1145/2815675.2815681; https://www.ftc.gov/system/files/documents/public_comments/2015/09/00011-97592.pdf).

Choey C., King, S. and Matsushima, N., 2016, *Pricing with cookies: behaviour-based price discrimination and spatial competition* (https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=EARIE43&paper_id=131&file_type=slides&utm_source=Bruegel+Updates&utm_campaign=656e7da39b-Blogs+review+11 %2F02 %2F2017&utm_medium=email&utm_term=0_eb026b984a-656e7da39b-278510293).

Choi,H., and Varian, H.V., 2011, 'Predicting the present with Google Trends', *The Economic Record*, Vol. 88, Special Issue, pp. 2-9.

Codagnone, C. and Martens, B., 2016, *Scoping the sharing economy: origins, definitions, impact, and regulatory issues*, Institute for Prospective Technological Studies, Digital Economy Working Paper 2016/0, Ispra, Italy.

Coles, M. G. and Smith, E., 1998, 'Market places and matching', *International Economic Review*, Vol. 39, No 1, pp. 239-254 (https://www.jstor.org/stable/pdf/2527239.pdf?seq=1#page_scan_tab_contents).

Connolly, K., 2016, 'Angela Merkel: Internet search Engines are "distorting perception"', *The Guardian*, 26 October (https://www.theguardian.com/world/2016/oct/27/angela-merkel-internet-search-engines-are-distorting-our-perception).

Couper, M. P. 2013, 'Is the sky falling? New technology, changing media, and the future of surveys'. *Survey Research Methods, 7: 145-56.*

Coyle, D., 2018, 'Do-it-yourself digital: the production boundary and the productivity puzzle', *Economica* (DOI:10.1111/ecca.12289; https://onlinelibrary.wiley.com/doi/10.1111/ecca.12289).

Coyle, D. and Nakamura, L., 2019, *Towards a framework for time use, welfare and household-centric economic measurement*) accessed 20 December 2018.

D'Andria, D., forthcoming, *The unbearable intangibility of the internet: taxing companies in the digital era*, JRC Science for Policy Brief.

De Hert, P. Papakonstantinou, V., Malgieri, G., Beslay, L. and Sanchez, I. 2018, 'The right to data portability in the GDPR: towards user-centric interoperability of digital services', *Computer Law and Security Review 2018*, pp. 193-203.

D'Onfro, J. and Browne, R., 2018 'EU fines Google $5 billion over Android antitrust abuse' CNBC (https://www.cnbc.com/2018/07/10/eu-hits-alphabet-google-with-android-antitrust-fine.html)

Dosis, A. and Sand-Zantman, W., 2018, *The ownership of data* (https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=IIOC2019&paper_id=433).

Duch-Brown, N., 2017a, *The competitive landscape of online platforms*, JRC Digital Economy Working Paper 2017-04.

Duch-Brown, N., 2017b, *Quality discrimination in online multi-sided markets*, JRC Digital Economy Working Paper 2017-06.

Duch-Brown, N., 2017c, *Platforms to business relations in online platform ecosystems*, JRC Digital Economy Working Paper 2017-07.

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

Economist, 2014, 'Fuel of the future: data is giving rise to a new economy', *The Economist*, 6 May (https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy).

Edelman, B., 2012, 'Using internet data for economic research', *Journal of Economic Perspectives*, Vol. 26, No 2, pp. 189-206.

Einav, L. and Levin, J. D., 2013, *The data revolution and economic analysis*, NBER Working Paper No 19035, May 2013 (https://www.nber.org/papers/w19035.pdf).

European Commission, 2016, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions — Online Platforms and the digital Single Market Opportunities and Challenges for Europe (COM(2016), 288 final, Brussels, 25.5.2016).

European Commission, 2017, 'Mergers: Commission fines Facebook €110 million for providing misleading information about WhatsApp takeover', European Commission press release (http://europa.eu/rapid/press-release_IP-17-1369_en.htm).

European Commission, 2018a, *Artificial intelligence: a European perspective* (https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective).

European Commission, 2018b, Proposal for a Council Directive lying down rules relating to the corporate taxation of a significant digital presence (COM(2018) 147 final, Brussels, 21.3.2018).

European Union, 2016, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5.2016, p. 1-88).

Evans, D. S., 2013, *Economics of vertical restraints for multi-sided platforms*, University of Chicago Institute for Law & Economics Olin Research Paper No 626 (http://ssrn.com/abstract=2195778).

Ezrachi, A. and Stucke, M. E., 2016, The rise of behavioural discrimination, Oxford Legal Studies Research Paper No 54/2016; University of Tennessee Legal Studies Research Paper (http://dx.doi.org/10.2139/ssrn.2830206).

Facebook (2014) *Annual report 2014* (http://www.annualreports.com/HostedData/AnnualReportArchive/f/NASDAQ_FB_2014.pdf).

Financial Times, 2019. Smart TVs sending private data to Netflix and Facebook. https://www.ft.com/content/23ab2f68-d957-11e9-8f9b-77216ebe1f17

Fradkin, A., 2015, *Search frictions and the design of online marketplaces* (https://pdfs.semanticscholar.org/b75a/56c4047b3df9d6ec84e49b24c6a2058346a6.pdf) accessed 26 September 2018.

FTC, 2014, *Data brokers: a call for transparency and accountability*, Federal Trade Commission (https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf).

Gerow, A. and Keane, M. T., 2012, *Mining the web for the 'voice of the herd' to track stock market bubbles* https://arxiv.org/ftp/arxiv/papers/1212/1212.2676.pdf

Grassegger, H. and Krogerus, M. 2017. The Data That Turned the World Upside Down: How Cambridge Analytica used your Facebook data to help the Donald Trump campaign in the 2016 election. https://www.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win

Gray, A., 2017, 'These are the world's 10 biggest corporate giants' World Economic Forum (https://www.weforum.org/agenda/2017/01/worlds-biggest-corporate-giants/).

Hall, J., and Krueger, A., 2015, *An analysis of the labor market for Uber's driver-partners in the United States*, Princeton University Working Paper 587 (http://dataspace.princeton.edu/jspui/bitstream/88435/dsp010z708z67d/5/587.pdf).

Hey, T., Stewart, T. and Tolle, K., 2009 *The fourth paradigm, data-intensive scientific discovery*, Microsoft Research, Redmond, WA.

Hirsch, L., 2018, ‘A year after Amazon announced its acquisition of Whole Foods, here’s where we stand’, CNBC, 15 June (https://www.cnbc.com/2018/06/15/a-year-after-amazon-announced-whole-foods-deal-heres-where-we-stand.html).

Hyrynsalmi, S., Mäkilä, T., Järvi, A., Suominen, A., Seppänen, M. and Knuutila, T., 2012, ‘App Store, marketplace, play! An analysis of multi-homing in mobile software ecosystems’, in Proceedings of the International Workshop on Software Ecosystems, IWSECO’2012, Cambridge, MA, pp. 55-68.

Hyrynsalmi, S., Seppänen, M. and Suominen, A., 2014, ‘Sources of value in application ecosystems’, *Information and Software Technology*, Vol. 56, pp. 1423-1435

Hyrynsalmi, S., Suominen, A. and. Mäntymäki. M., 2016, ‘The influence of developer multi-homing on competition between software ecosystems’, *Journal of Systems and Software*, Vol. 111, pp. 119-127.

Investopedia, 2015, ‘Google’s revenue beats the GDP of several major countries’ (https://www.investopedia.com/articles/investing/061115/googles-revenue-beats-gdp-several-major-countries.asp).

Isaac, M., 2017, ‘How Uber deceives the authorities worldwide’, *New York Times*, 3 March (https://www.nytimes.com/2017/03/03/technology/uber-greyball-program-evade-authorities.html).

Janetzko, D., 2014, ‘Predictive modeling in turbulent times — what Twitter reveals about the EUR/USD exchange rate’, *NETNOMICS Economic Research and Electronic Networking*, Vol. 15, No 2 (DOI: 10.1007/s11066-014-9087-y).

Jones, C. I. and Tonetti, C., 2018, *Nonrivalry and the economics of data*, Version 0.6, 31 July (https://www.gsb.stanford.edu/faculty-research/working-papers/nonrivalry-economics-data).

Kaminska, I., 2019, ‘The taxi unicorn’s new clothes’, *Financial Times*, 1 December (https://ftalphaville.ft.com/2016/12/01/2180647/the-taxi-unicorns-new-clothes).

Kee, T. H., 2018, ‘Trump has trained stock market investors’, Market Watch (https://www.marketwatch.com/story/trump-has-trained-stock-market-investors-2018-07-20) accessed October 2018.

Khan, M., and Brunsden, J., 2018, ‘France and Germany abandon plans for EU digital tax’, *Financial Times*, 4 December (https://www.ft.com/content/fc7330d4-f730-11e8-af46-2022a0b02a6c).

Kitchin, R., 2014, ‘Big Data, new epistemologies and paradigm shifts’, *Big Data & Society*, April-June, pp. 1-12. (DOI: 10.1177/2053951714528481bds.sagepub.com).

Kosinski, M., Stillwella, D. and Graepelb, T., 2013, ‘Private traits and attributes are predictable from digital records of human behaviour’, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 110, No 15, pp. 5802-5805 (http://www.pnas.org/content/early/2013/03/06/1218772110.full.pdf+html).

Kramer, A. and Kalka, R., 2016, ‘How digital disruption changes pricing strategies and price models’, in Khare, A., Stewart, B. and Schatz, R. (eds), *Phantom ex machina: digital disruption’s role in business model transformation*, Springer, Dordrecht (https://doi.org/10.1007/978-3-319-44468-0).

Krugman, P., 2014, ‘Amazon’s monopsony is not O.K.’, *The New York Times*, 19 October (https://www.nytimes.com/2014/10/20/opinion/paul-krugman-amazons-monopsony-is-not-ok.html).

Laney, D., 2001, ‘3D data management controlling data volume, velocity, and variety’, META Group (http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf).

Laney, D., 2012, Deja Vvvu: others claiming Gartner’s construct for big data (http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/).

Lazer, D., Kennedy, R., King, G. and Vespignani, A., 2014, ‘The parable of Google flu: traps in big data analysis’, *Science*, Vol. 343, 14 March.

Liem, C., and Petropoulos, G., 2016, ‘The economic value of personal data for online platforms, firms and consumers’, Bruegel blogspot, 14 January (http://bruegel.org/2016/01/the-economic-value-of-personal-data-for-online-platforms-firms-and-consumers/).

Malhotra, A. and Van Alstyne, M., 2014, ‘The dark side of the sharing economy … and how to lighten it’, *Communications of the ACM*, Vol. 57, pp. 24-27.

Macione, V. 2019. Facebook diventa uno Stato e nasce il tribunale di Zuckerberg. Il Giornale 19/09/2019. http://www.ilgiornale.it/news/cronache/svolta-social-network-1755158.html

Martens, B., 2016, *An economic policy perspective on online platforms*, Institute for Prospective Technical Studies Digital Economy Working Paper 2016/05.

Mayer-Schonberger, V. and Cukier, K., 2013, *Big data: a revolution that will transform how we live, work and think*, Houghton Mifflin Harcourt, Boston, MA.

McArdle, M., 2019, ‘Uber and Lyft are losing money. At some point, we’ll pay for it’, *Washington Post*, 5 March (https://www.washingtonpost.com/opinions/uber-and-lyft-are-losing-money-at-some-point-well-pay-for-it/2019/03/05/addd607c-3f95-11e9-a0d3-1210e58a94cf_story.html).

Mckinsey, 2018, *Superstars. The dynamics of firms, sectors, and cities leading the global economy*, Mckinsey Global Institute Discussion Paper, October 2018. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/innovation/superstars%20the%20dynamics%20of%20firms%20sectors%20and%20cities%20leading%20the%20global%20economy/mgi_superstars_discussion%20paper_oct%202018-v2.ashx

Mckinsey, 2019, *Twenty-five years of digitization: ten insights into how to play it right*, Briefing note prepared for the Digital Enterprise Show, Madrid, 21-23 May, Mckinsey Global Institute (https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/twenty-five-years-of-digitization-ten-insights-into-how-to-play-it-right).

Michael, R. and Stiglitz, J., 1976, ‘Equilibrium in competitive insurance markets: an essay on the economics of imperfect information’, *Quarterly Journal of Economics*, Vol. 90, No 4, pp. 629-649 (http://www.jstor.org/stable/1885326).

Mikians, J., Gyarmati, L., Erramilli, V. and Nikolaos Laoutaris, N., 2012, *Detecting price and search discrimination on the Internet* (http://conferences.sigcomm.org/hotnets/2012/papers/hotnets12-final94.pdf?utm_source=Bruegel+Updates&utm_campaign=656e7da39b-Blogs+review+11 %2F02 %2F2017&utm_medium=email&utm_term=0_eb026b984a-656e7da39b-278510293).

Möhlmann, M. and Zalmanson, L., 2017, ‘Hands on the wheel: navigating algorithmic management and Uber drivers’ autonomy’, in Proceedings of the International Conference on Information Systems (ICIS 2017), 10-13 December, Seoul, South Korea.

Munoz de Bustillo, R., 2019, Key Challenges for the European Welfare States. JRC Working Papers Series on Labour, Education and Technology.

Musgrave, R. A., 1959, The theory of public finance: a study in public economy, McGraw-Hill, New York.

OECD, 2015, *Data driven Innovation. Big data for growth and well-being*, OECD Publishing, Paris (https://www.oecd.org/sti/data-driven-innovation-9789264229358-en.htm).

OECD, 2019, *Digital innovation: seizing policy opportunities*, Organisation for Economic Co-operation and Development http://www.oecd.org/publications/digital-innovation-a298dc87-en.htm

Palfrey, J. and Gasser, U., 2012, *Interop: the promise and perils of highly interconnected systems* (https://cyber.harvard.edu/publications/2012/interop).

Pawelke, A. and Tatevossian, A. R., 2013, ‘Data philanthropy: where are we now?’, United Nations Global Pulse, 8 May (https://www.unglobalpulse.org/data-philanthropy-where-are-we-now).

Pedraza, P. de, Visitin, S., Tijdens, K. and Kismihok, G., 2019, ‘Survey vs scraped data: comparing time series properties of web and survey vacancy data’, *IZA Journal of Labour Economics* 8:4. https://doi.org/10.2478/izajole-2019-0004

Pilaszy, I. and D. Tikk, 2009, ‘Recommending movies: even a few data is more valuable than metadata’, in Proceedings of the 2009 ACM Conference on Recommender Systems, pp. 93-100 (DOI: 10.1145/1639714.1639731).

Posner, E. and Weyl, G., 2018, *Radical markets: uprooting capitalism and democracy for a just society*, Princeton University Press, Princeton, NJ.

Pratley, N., 2018, 'UK finally takes on arrogant tech giants with digital services tax. Budget levy on giants such as Facebook, Google and Amazon could go further — but it's a start', *The Guardian*, 29 October (https://www.theguardian.com/uk-news/2018/oct/29/uk-digital-services-tax-budget-facebook-google-amazon).

Prewitt, K., 2013, 'The 2012 Morris Hansen lecture: thank you Morris, et al., for Westat, et al.', *Journal of Official Statistics*, Vol. 29, No 2, pp. 223-231.

Reips, U.-D. and Garaizar, P., 2011, 'Mining Twitter: microblogging as a source for psychological wisdom of the crowds', *Behavior Research Methods*, Vol. 43, pp. 635-642 (doi:10.3758/s13428-011-0116-6).

Rubens, P., 2014, 'Can cloud storage costs fall to zero? Cloud storage providers keep lowering their prices. How low can they go?', Enterprise Storage, 6 August (https://www.enterprisestorageforum.com/storage-management/can-cloud-storage-costs-fall-to-zero-1.html) accessed 10 February 2019.

Rysman, M. (2009). The Economics of Two-Sided Markets, *Journal of Economic Perspectives*, vol. 23, No. 3, Summer 2009, pp. 125-43. https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.23.3.125

Samuelson, P. A., 1948, *Economics*, McGraw-Hill, New York.

Samuelson, P. A., 1954, 'The pure theory of public expenditure', *Review of Economics and Statistics*, Vol. 36, No 4, pp. 387-389 (doi:10.2307/1925895).

Samuelson, P. A., 1955, 'Diagrammatic exposition of a theory of public expenditure', *Review of Economics and Statistics*, Vol. 37, No 4, pp. 350-356 (doi:10.2307/1925849).

Samuelson, P. A. and Nordhaus, W. D, 2010, *Economics*, 19th ed, McGraw-Hill, New York.

Sandle, P., 2018, Britain to target online giants with new 'Digital Services Tax', Reuters, 29 October. (https://uk.reuters.com/article/us-britain-budget-digital-tax/britain-to-target-online-giants-with-new-digital-services-tax-idUKKCN1N3265).

Schroeder, R. and Cowls, J., 2014, *Big data, ethics, and the social implications of knowledge production* (https://dataethics.github.io/proceedings/BigDataEthicsandtheSocialImplicationsofKnowledgeProduction.pdf).

Scott, M. and Young, Z., 2018, 'France and Facebook announce partnership against online hate speech. Emmanuel Macron has teamed up with Mark Zuckerberg to review the country's regulatory response to the issue', Politico, 11 December (https://www.politico.eu/article/emmanuel-macron-mark-zuckberg-paris-hate-speech-igf/).

Scott Morton, F., Bouvier, P., Ezracchi, A., Jullien, B., Kazt, R. Kimmelman, G., Melamed, A. D. and Morgenstern, J., 2019, *Report for the study of digital platforms market structure and anti-trust subcommittee*, 15 May. https://www.judiciary.senate.gov/imo/media/doc/market-structure-report%20-15-may-2019.pdf

Shiller, B. R., 2014, *First-degree price discrimination using big data* (http://benjaminshiller.com/images/First_Degree_PD_Using_Big_Data_Jan_27,_2014.pdf?utm_source=Bruegel+Updates&utm_campaign=656e7da39b-Blogs+review+11 %2F02 %2F2017&utm_medium=email&utm_term=0_eb026b984a-656e7da39b-278510293).

Smith, A. (1776). An Inquiry into the Nature and causes of the wealth of nations. London: A. Strahan, T. Cadell. https://books.google.it/books?id=PAQMAAAAYAAJ&dq=editions:HkmbBXCA1kcC&pg=PR1&redir_esc=y#v=onepage&q&f=true

Statista (2015) 'Google's annualized advertising ARPU from the 1st quarter of 2012 to the 1st quarter of 2014 (in US dollars)' (http://www.statista.com/statistics/306570/google-annualized-advertising-arpu/) accessed 7 December 2015.

Steel, E., 2013, 'Companies scramble for consumer data', *Financial Times*, 12 June (https://www.ft.com/content/f0b6edc0-d342-11e2-b3ff-00144feab7de) accessed 28 November 2018.

Steel, E., Locke, C., Cadman, E. and Freese, B., 2013, 'How much is your personal data worth? Use our calculator to check how much multibillion-dollar data broker industry might pay for your personal data', *Financial Times*, 12 June (https://ig.ft.com/how-much-is-your-personal-data-worth/) accessed 28 November 2018.

Stiglitz, J. P., 2001, 'Information and change in the paradigm in economics', Prize lecture, 8 December (https://www.nobelprize.org/prizes/economic-sciences/2001/stiglitz/lecture/).

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. and Hassabis, D., 2017, 'Mastering the game of Go without human knowledge', *Nature*, Vol. 550 (doi:10.1038/nature24270; https://www.nature.com/articles/nature24270.pdf).

Taylor, L., Schroeder R. and Meyer, E., 2014, 'Emerging practices and perspectives on big data analysis in economics: bigger and better or more of the same?', *Big Data & Society* July-December, pp. 1-10.

Tett, G., 2018, 'Recalculating GDP for the Facebook age. The true impact of social media? Economists are approaching the question from a different angle', *Financial Times* (https://www.ft.com/content/93ffec82-ed2a-11e8-8180-9cf212677a57).

Uber, 2018, 'Fraudulent trips: how to recognise fraud' (https://www.uber.com/en-ZA/drive/resources/recognising-fraud/) accessed January 2019.

UN, 2013, *A new global partnership: eradicate poverty and transform economies through sustainable development*, United Nations. https://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=893&menu=1561.

UN, 2014, *A world that counts: mobilising the data revolution for sustainable development*, Independent Experts Advisory Group on Data Revolution for Sustainable Development, November 2014, United Nations (http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf).

Unctad, 2017, *World investment report 2017, investment and the digital economy*, UN Publications, Geneva.

Ulloa, J., 2019, 'Newsom wants companies collecting personal data to share the wealth with Californians', *Los Angeles Times*, 5 May (https://www.latimes.com/politics/la-pol-ca-gavin-newsom-california-data-dividend-20190505-story.html).

Ursu, R., 2015, *The power of rankings: quantifying the effect of rankings on online consumer search and purchase decisions* (https://pdfs.semanticscholar.org/b833/9f0aa1d919fdd7b58a533174eaf463645403.pdf) accessed 26 September 2018.

UTI, 2018, *Assessing the impact of artificial intelligence*, International Telecommunications Union Issue Paper No 1, September 2018 (https://www.itu.int/pub/S-GEN-ISSUEPAPER-2018-1).

Varian, H. R., 2013, 'Big data: new tricks for econometrics', *Journal of Economic Perspectives*, Vol. 28, No 2, pp. 3-28.

Vaughan, R. and Hawksworth, J., 2014, *The sharing economy: how will it disrupt your business? Megatrends: the collisions*, PriceWaterhouseCoopers, London.

Vigo, R., 2013, 'Complexity over uncertainty in generalized representational information theory (GRIT): a structure-sensitive general theory of information', *Information*, Vol. 4, pp. 1-30 (http://cogprints.org/8784/1/Vigo%20(2013).pdf).

Vitali S., Glattfelder J. B. and Battiston, S., 2011, 'The network of global corporate control', *PLoS ONE* Vol. 6, No 10, pp. e25995 (doi:10.1371/journal.pone.0025995).

Vogel, C. and Janssen, J., 2009, 'Emoticonsciousness', in Esposito, A., Hussain, A., Marinaro, M. and Martone, R. (eds), *Multimodal signals: cognitive and algorithmic issues*, Lecture Notes in Computer Science, Vol. 5398/2009, Springer, Dordrecht, pp. 271-287.

WEF, 2013, *Young global leaders sharing economy*, World Economic Forum Dialogue Position Paper (http://www3.weforum.org/docs/WEF_YGL_CircularEconomyInnovation_PositionPaper_2013.pdf).

WEF, 2014, Towards the circular economy: accelerating the scale-up across global supply chains, World Economic Forum (http://www3.weforum.org/docs/WEF_ENV_TowardsCircularEconomy_Report_2014.pdf).

White House, 2015, Big data and differential pricing, White House Council of Economic Advisers (https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf?utm_source=Bruegel+Updates&utm_campaign=656e7da39b-Blogs+review+11 %2F02 %2F2017&utm_medium=email&utm_term=0_eb026b984a-656e7da39b-278510293).

Wong, J. C., 2017, 'Greyball: how Uber used secret software to dodge the law', *The Guardian*, 4 March (https://www.theguardian.com/technology/2017/mar/03/uber-secret-program-greyball-resignation-ed-baker).

Worstall, T., 2017, 'Understanding the economic value of your personal data'. www.computerweekly.com 26 May 2016 (https://www.computerweekly.com/opinion/Understanding-the-economic-value-of-your-personal-data)

List of abbreviations and definitions

AI	artificial intelligence
BD	big data
GDPR	general data protection regulation

List of figures

List of tables

KJ-NA-29825-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office
of the European Union

doi:10.2760/668

ISBN 978-92-76-09231-5