Martens, Bertin

Working Paper

The impact of data access regimes on artificial intelligence and machine learning

JRC Digital Economy Working Paper, No. 2018-09

Provided in Cooperation with:
Joint Research Centre (JRC), European Commission

Suggested Citation: Martens, Bertin (2018) : The impact of data access regimes on artificial intelligence and machine learning, JRC Digital Economy Working Paper, No. 2018-09, European Commission, Joint Research Centre (JRC), Seville

This Version is available at:
https://hdl.handle.net/10419/202237

European Commission

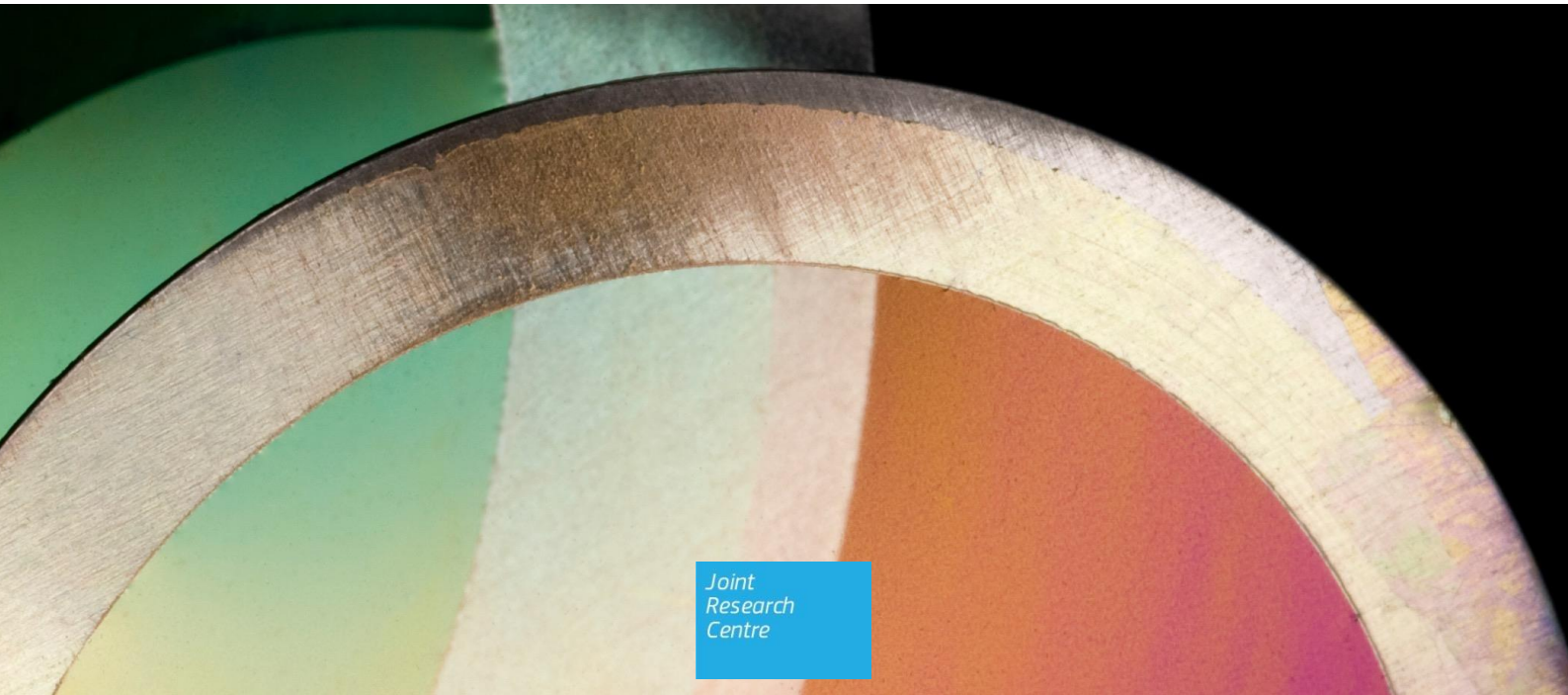
JRC TECHNICAL REPORTS

JRC Digital Economy Working Paper 2018-09

The impact of data access regimes on artificial intelligence and machine learning

Bertin Martens

December 2018



This publication is a Working Paper by the Joint Research Centre, the European Commission's in-house science service. It aims to provide evidence-based scientific support to the European policy-making process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

Contact information
European Commission, Joint Research Centre
Address: Edificio Expo. c/Inca Garcilaso, 3. 41092 Seville (Spain)
E-mail: bertin.martens@ec.europa.eu
Tel.: +34 954488318

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

JRC114990

ISSN 1831-9408 (online)

Seville, Spain: European Commission, 2018

How to cite: Bertin Martens ; The impact of data access regimes on artificial intelligence and machine learning , Digital Economy Working Paper 2018-09; JRC Technical Reports.

Table of Contents

Abstract

Digitization triggered a steep drop in the cost of information. The resulting data glut created a bottleneck because human cognitive capacity is unable to cope with large amounts of information. Artificial intelligence and machine learning (AI/ML) triggered a similar drop in the cost of machine-based decision-making and helps in overcoming this bottleneck. Substantial change in the relative price of resources puts pressure on ownership and access rights to these resources. This explains pressure on access rights to data. ML thrives on access to big and varied datasets. We discuss the implications of access regimes for the development of AI in its current form of ML. The economic characteristics of data (non-rivalry, economies of scale and scope) favour data aggregation in big datasets. Non-rivalry implies the need for exclusive rights in order to incentivise data production when it is costly. The balance between access and exclusion is at the centre of the debate on data regimes. We explore the economic implications of several modalities for access to data, ranging from exclusive monopolistic control to monopolistic competition and free access. Regulatory intervention may push the market beyond voluntary exchanges, either towards more openness or reduced access. This may generate private costs for firms and individuals. Society can choose to do so if the social benefits of this intervention outweigh the private costs. We briefly discuss the main EU legal instruments that are relevant for data access and ownership, including the General Data Protection Regulation (GDPR) that defines the rights of data subjects with respect to their personal data and the Database Directive (DBD) that grants ownership rights to database producers. These two instruments leave a wide legal no-man's land where data access is ruled by bilateral contracts and Technical Protection Measures that give exclusive control to de facto data holders, and by market forces that drive access, trade and pricing of data. The absence of exclusive rights might facilitate data sharing and access or it may result in a segmented data landscape where data aggregation for ML purposes is hard to achieve. It is unclear if incompletely specified ownership and access rights maximize the welfare of society and facilitate the development of AI/ML.

This is a background paper to the JRC report "Artificial Intelligence: A European perspective" (2018).

1. Introduction

Over the last decade Machine Learning (ML) emerged as the dominant Artificial Intelligence (AI) technology. It could be considered as a vastly scaled-up version of existing statistical analysis techniques (Igami, 2018) that facilitate the extraction of insights from large and complex datasets and use these for prediction and making choices in a more efficient way (Agarwal et al, 2018). From an economic perspective ML is a technology that reduces the cost and increases the efficiency of decision making in complex environments. ML algorithms require a "ground truth" training dataset that contains input data and desired outcomes. The learning component consists of gradually connecting the inputs to the desired outputs. Once the algorithm has reached a sufficient degree of accuracy it can apply this learning to new input data in order to produce the desired decision outputs. Some spectacular results have been achieved in the last couple of years whereby ML algorithms have become better than humans in taking decisions, for example in games like chess and go.

Scholars of AI have pointed out that AI/ML thrives on access to (big) datasets as necessary inputs for training algorithms. Data ownership and access rules matter. Cockburn et al (2018) suggest that ML has changed the production process for innovative ideas that has become much more data driven. Ensuring access to data becomes an important issue for promoting innovation. "If there are increasing returns to scale or scope in data acquisition (there is more learning to be had from the “larger” dataset), it is possible that early or aggressive entrants into a particular application area may be able to create a substantial and long-lasting competitive advantage over potential rivals merely through the control over data rather than through formal intellectual property or demand side network effects". ML techniques put a premium value on larger datasets. The non-rival nature of data adds another argument in favour of data sharing, "open data" and more generally facilitating access to data. The OECD (2015) argues that data should be considered as a public rather than a private good and data should be under an open access regime that permits wide sharing in order to promote data-driven innovation.

A more balanced economic approach would reformulate the argument in terms of private and social value of data. That starts from the private interests and incentives of the data supplier and the data collector or processor. The private costs and benefits can differ from the social value of the data for society as a whole. When free data markets create a gap between private and social value a case may be made for market failure and regulatory intervention. Regulators may favour more data openness, or possibly the opposite and facilitate exclusive data access rights. In the EU, data regulation goes in both directions: some rules make data access more exclusive while others reduce exclusivity. That bi-directional approach may sound paradoxical but it is not necessarily a contradiction. Intellectual property rights regimes already combine elements of exclusiveness and openness, and data access regimes could follow a similar economic logic in order to find an appropriate balance between private and societal interests. In practice however there are few legal instruments regarding ownership or access rights to data. The well-known EU General Data Protection Regulation (GDPR) covers rights related to personal data and the EU Database Directive establishes exclusive ownership rights for "databases", subject to some restrictive conditions. There are also specific sector and

thematic regulatory instruments that affect access to particular types of datasets. In many cases however data holders create a de facto ownership situation by means of technical protection measures (TPMs) that restrict access to the data they collected. This results in a market-based data access regime that enables contractual and market-based trade in data and data-based services. While this setting may generate private benefits it remains to be seen if it is socially efficient for society as a whole. The rise of AI and ML technologies adds another layer of questions to this setting: does the data access regime promote AI-driven technological innovation?

This paper takes a closer look at market-based data access regimes. These may range from free and open access to conditional and paid access or to no access at all. For paid access, supply and demand and market structure play an important role. Data can be traded directly between two parties or can be monetised as indirect data services. In the latter case there is no transfer of data, only the delivery of a data-based service. Perhaps more importantly, direct and indirect data trade often have an impact on markets for "real" goods and services. Indeed, data and data services are not traded for their own sake but because they are an essential input for deciding on transactions in other markets. Data transactions are an upstream market that affects downstream markets for goods and services. The characteristic of the upstream intermediary input market (monopolistic, competitive) may have implications for the downstream final decision market. We examine the economic characteristics of different types of data, the market structure that they may induce and their potential implications for downstream markets. We can then examine the impact of possible regulatory intervention to open up or reduce data market access, and how this could affect the development of ML and AI.

This paper is structured as follows. Section 2 discusses the economic relationship between data markets and the development of AI/ML. Section 3 investigates the economic characteristics of data and why this leads to a paradox that favours at the same time wider access to data but also exclusive control over data. Section 4 explores a number of scenarios regarding access to data in a free data market setting, without regulatory intervention and with technical protection measures as the only access barrier. Section 5 describes the current regulatory setting for data transactions in the EU and how it includes a combination of factors that promote as well as inhibit data sharing. Section 6 concludes.

2. The economic impact of data and AI/ML

Prior to digitization information came in many analogue formats and carriers. Text and figures were written on paper, sound recorded on a magnetic tape, pictures on a silver-coated transparent plastic tape, etc. These formats were not compatible or interoperable. It was costly to store, process and communicate them. These high costs diminished the net benefits of data and made it worthwhile to collect them only for very valuable applications. For example: journalists stalking famous people or detective and criminal investigations. In that pre-digital setting there was not much need to define exclusive data or information ownership and access rights since personal and company data were difficult and costly to obtain. Some minimal privacy rights were sufficient to cover these situations.

Digital technology changed all that[1]. The introduction of a universal digital (binary 0/1) data format with a standardised electronic carrier has greatly reduced the costs of interoperability at the level of the information carrier[2]. It can easily be captured, stored, processed and transmitted at very low cost. It triggered a massive drop in the cost of information. This price effect in turn led to quantity and variety effects – the production of a huge amount of data, or "big data" – and a substitution effect – shifting activities away from non-digital to digital information environments. For example, consumers shifted their purchases from offline bricks & mortar stores to online digital stores where they can choose from a much wider variety of goods and services. This increases competition between sellers and puts downward pressure on the prices of goods and services. That triggers positive and negative welfare effects across society (Martens, 2013; Cardona et al, 2017).

The sudden availability of vast amounts of digital data at very low cost has major implication for goods and services markets, and for human social interaction. Data, or information, are a necessary intermediate input into making individual choices and carrying out transactions (in goods, services, or data) between people. They have no economic value as such but acquire value when they contribute to making a more informed choice or transaction. Put differently, there is a vertical relationship between upstream data markets and downstream markets for transactions in goods and services: changes in access and conditions in the upstream market will affect the downstream market. For example, access to digital car data markets affects aftersales services markets for cars (Martens and Mueller-Langer, 2018).

The benefits derived from these choices and transaction will determine the value of the data that were required to do it. These benefits may be unevenly distributed between the parties in a transaction, because of differences in market power and information asymmetries between the parties. Data may thus have different values for different parties. Prior to digital data we already made choices and carried out transactions of course. The introduction of digital data may change choices and transaction patterns and thereby trigger shifts in benefits or in the economic "value chain": who gains, who loses? The additional value derived from transactions driven by digital data, compared to pre-digital information, determines the value of digital data. The label "digital transformation" is often used to refer to these changes in transaction patterns and benefits induced by digital data technologies.

The shift towards online activities has created a huge "commons" of easily accessible data that are cheap to collect and store in privately controlled servers whereby access is protected by technical measures, not by legal ownership rights. The ubiquity of digital devices that collect indelible traces of people's behaviour make it easy to collect personal data and cause disputes about privacy and access to personal data. It also facilitates access to many types of firm data and activities because

[1] See Duch-Brown et al, 2017, chapter 2 for a more detailed discussion of the impact of digital information technology.

[2] Achieving full interoperability involves more steps and levels (technical, semantic, organizational interoperability, etc). See for example chapter 3 in the European Interoperability Framework (https://ec.europa.eu/isa2/eif_en). Even at the technical level alone it involves achieving interoperability between applications and infrastructures, linking systems and interconnection services, etc.

they leave digital traces in the online "commons" pool. Privatisation of the data commons has resulted in an "anti-commons": many useful data are hard to access and potential benefits remain unexploited. That is not optimal from a societal perspective. It is a "might makes right" regime (Umbeck, 1981) whereby access and use depends on the economic power to collect data and store them in protected servers from where they can used to produce data-driven services. There are many cases where market-based contracts lead to distorted and suboptimal situations (Duch-Brown et al, 2017, chapter 4). Case-by-case data access regulation may solve some of these bottlenecks but it may also lead to a fragmented and complex data regime with new distortions due to regulatory capture by special interest groups. Predictably, this anti-commons data access regime has triggered a counter-movement: the push for more data sharing, free and open access, etc. But full and open access to personal and commercial data is not socially desirable (Palfrey and Grasser, 2012). It would seriously undermine the privacy of individuals and the commercial secrecy of firms. That explains the clamour for new data protection laws and ownership rights that would make access illegal under certain conditions. Stakeholders fear that wider access would disrupt their business models and create new "data value chains". This emerging data access debate is not surprising. Institutional economists have argued long ago that changes in relative resource prices put pressure on the institutions and property rights that regulate access to these resources (North, 1994). How to find a balance between open and closed data access has become a big societal debate that is far from finished.

Apart from access rights, the sudden data glut also created new problems for human intelligence. Human information processing capacity is unable to cope with large quantities of information. In the "attention economy" we have to fragment our limited attention span and processing capacity over a large array of information inputs in an attempt to select the most relevant signals. A first solution for this bottleneck was filtering or pre-selection of the most relevant information. Search engines and recommender systems for example produce a pre-selected set of potentially relevant items to reduce search costs and facilitate human choice. But search is not sufficient in many cases. ML is a new tool that helps us wade through oceans of data and find the "needle in the haystack" (Agarwal et al, 2018) that we are looking for. ML creates a second wave in data cost reduction by decreasing the cost of data-based decision making. That, in turn, triggers a similar debate on access to data and algorithms that are inputs into human and automated decision-making. There are concerns that ML will replace human decisions in work situation and generate widespread unemployment (Martens and Tolan, 2018). However ML may also assist humans in their day-to-day consumer decisions. ML boosts the benefits that could be extracted from very large datasets and explains the data "gold rush", the quest to get access to and collect ever more data.

3. The economic characteristics of data

In this section we discuss three economic characteristics of data analytics and decision-making technologies: economies of scale, economies of scope and non-rivalry.

2.1. Economies of scale

ML is a statistical technique. Statistical estimates become more reliable as the size of the underlying dataset increases and the variation in the data decreases. With low variety, results will soon converge to a robust outcome. With high variety, larger sample datasets are required for robustness. With high variety in the observations and many variables in the dataset, achieving robust predictions may require very large datasets. Economies of scale in the size of datasets may be subject to diminishing returns. Scattered empirical evidence suggests that in some cases diminishing returns may set in at a very early stage (Pilaszy & Tikk, 2009, on film selection) while in other cases it only arrives when the number of observations increases many orders of magnitude (Varian, 2014) or never (Lewis & Rao, 2015, on the efficiency of online advertising).

ML is not only *able* to handle much large datasets than human can handle; it also *requires* much larger datasets than humans need in order to learn. While humans may only require a few observations to learn a behavioural response, ML may require thousands or millions of observations to learn some basic responses. For example, a computer could beat a human Go player after learning strategies from playing and observing millions of games. The human challenger had "only" played 50,000 games in his lifetime to reach that level. A self-driving car algorithm can handle most traffic situations after having "learned" to drive from millions of kilometres of data inputs; a human driver only needs a few thousand kms of experience to become a proficient driver. Still, the algorithm can digest these millions of kms of data input in a much shorter time than a human driver needs to drive a few thousand kms. Moreover the algorithm can drive many cars at the same time; a human driver can only drive one car at the time.

Investing in high quality datasets for training of ML algorithms is costly and implies high fixed costs. Once there are trained however the marginal cost of additional use of the algorithm can be very low. That gives rise to economies of scale in ML.

2.2. Economies of scope

There may also be economies of scope (Rosen, 1983) in merging two or more datasets. Economies of scope occur when the benefits of analyzing/using a joint dataset are higher than the sum of benefits of analyzing/using each dataset separately: $V(d1,d2) > V(d1) + V(d2)$. These benefits occur when there is a relationship between the two sets, i.e. when they are not completely separable and data that pertain to one situation may also be relevant for another situation. For example, web surfing data may produce insights on consumer behaviour. Merging these with mobile phone data may produce more insights, compared to studying both datasets separately. Adding pay data in shops adds further insights, etc. Applying ML separately to each of these datasets would not produce the same complex insights as applying it to the aggregated set. By contrast, there is no point in merging two datasets for e-commerce and playing chess because they are fully separable. A single ML algorithm would not gain from learning from that aggregated set. On the contrary, it would only confuse the algorithm. Economies of scope explain the business model of data aggregators who combine data from various sources into a single consistent dataset: the wider the coverage of the dataset in terms of variety of situations and observations, the more accurate and insightful the predictions that can be made with these datasets.

Data aggregation is not always useful however. For example, individual car data on driving behaviour and the mechanical state of the car, collected by digitally connected cars, are valuable for insurance and maintenance purposes. There is no need to aggregate them with data from other cars in order to increase the value of the data. By contrast, car navigation data need to be aggregated by a navigation service provider in order to identify traffic jams and send this information back to drivers. There are substantial economies of scope in the aggregation compared to the marginal value of each individual car navigation dataset (Bergemann and Bonatti, 2018). Similarly, large e-commerce platforms can aggregate data across many transactions that give them a comprehensive market overview that is more valuable than separate data on individual transactions.

Economies of scope drive a wedge between a relatively low private and higher collective or social value of datasets. They add another argument in favour of access to data. They change the balance between exclusive protection and wider access. Exclusive access results in monopolies and static price inefficiency which is often considered to be the price to be paid in order to achieve dynamic innovation efficiency. However, with a premium on data pooling and less exclusive access, less protection of access rights would facilitate wider access to data, establishing data pools and promote innovation. Not the data but the production of insights from data is what needs to be incentivised in that case.

Economies of scope can explain "barter" trade in data, or the exchange of data in return for "free" services", an ubiquitous online business model. For example, consumers are willing to share personal data with a search engine or a social media site in return for an information service or a social contacts service. They have an economic incentive to do so because the market value of individual personal data is very low compared to the consumer surplus value of the free service they receive in return from the data aggregator (Brynjolfsson, Aganamimi and Eggerts, 2018). Search and social media platforms aggregate these data and that produces a more valuable market overview that can be monetized, for example via advertising. Individual users cannot realize that value by keeping their data separate.

Data aggregation is a necessity for ML algorithms because they require large datasets. Moreover, the capacity of algorithms to digest large datasets makes them useful to handle “needles in haystacks” problems that humans cannot deal with: complex decisions that occur only very rarely and require a large dataset (Agarwal et al, 2018)[3]. Common machine learning methods do not perform well with small data sample sizes, especially when data are noisy. Machine learning models tend to 'overfit' in these cases, reducing the efficiency of algorithms once they are applied in real-life settings. One way to address this is to try to understand the underlying mechanisms in the data through causal inference. This potentially increases the interpretability and fairness of algorithms (Athey, 2017). Causal interference is what human learners do when they encounter a novel situation: they try to understand

[3] Kai-Fun Lee (2018) points out that China has an advantage in data collection, not only because of its large population but also because there is more intense use of (mobile) internet and therefore generates more data. That big data supply advantage could result in an advantage in AI/ML because more data are available for training the algorithms and producing more valuable insights.

how it arises and how to respond to it. Contrary to current ML, humans usually do not need to experience a novel situation many times in order to understand it.

2.3. Non-rivalry

A third economic characteristic of data that plays a role in the debate on open access and data sharing is non-rivalry. Goods and services are rival products: they can only be used by one person at the same time. If several persons want to read the same book at the same time, drive the same car or eat the same meal, this creates interference and costs for all involved. An idea, an invention, a piece of information is non-rival: it can be used by many people at the same time. As a result, the benefits derived from data are higher when used by a group of people compared to when it is used by a single individual. From a societal perspective, it may therefore be better to share data as widely as possible rather than keeping them private. This is the "public good" argument that underpins the OECD (2015) report.

However, the flip side is that wide sharing diminishes their economic value. This may reduce incentives to invest in the production of data. This problem was first discovered by authors and publishers of books and by inventors of new technologies. All these innovative ideas and artworks could easily be copied by others, leaving the original creator without any remuneration for his efforts. To avoid this, non-rival products need to be made excludable by laws that grant exclusive rights to an owner. In the case of intellectual property this exclusive right is given to the creators or inventors.

Non-excludability may not be a problem when data are a by-product of on-going activities that require no additional incentives. For instance, e-commerce firms store consumer behaviour data on their websites as part of their online transactions (Kerber, 2017). Public administrations collect loads of data while carrying out their duties. In other cases firms may deliberately invest in collecting data by installing sensors and setting up servers to collect, store, transmit and process the data. Without the ability to monetise revenue from the data there would be no incentive to collect them.

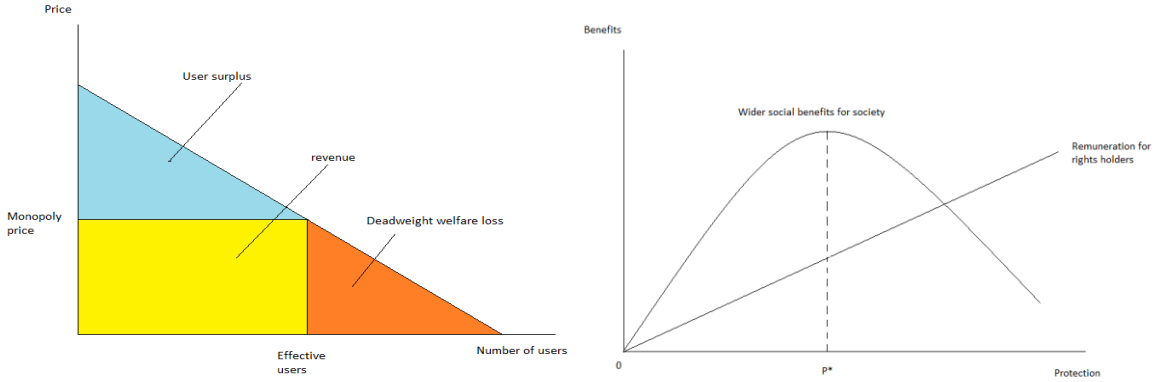
2.4. The trade-off in exclusive ownership rights

Non-rivalry leads to a paradox: wide sharing of non-rival goods is beneficial for society but society gives exclusive ownership rights to the producers of non-rival goods that enable them to prevent unauthorised and non-remunerated sharing. This paradox is explained by the economic theory of intellectual property rights (IPR) that compares the drawbacks from a private monopoly use right with the dynamic benefits from granting such a right. One the one hand IP rights give a monopoly right to a private party that gets exclusive rights over the use of the IP-protected item. This monopoly is socially inefficient because it generates deadweight welfare losses: not all users who could use it get access to it.

Figure 1a shows the economic consequences of attributing an exclusive ownership right to the producer of an idea. The owner can, if he wishes to do so, sell the information at a monopoly price (by means of licenses for example) that will attract a number of buyers or users. The owner will gain

revenue (the yellow rectangle). Consumers who were prepared to pay an even higher price will benefit from a "consumer surplus" - the difference between what they were willing to pay and what they actually paid (the blue triangle). But not everybody will be willing to pay that price. Some potential users are left out of the market (the orange triangle). We call this the social deadweight losses from monopoly: nobody benefits from these losses, neither consumers nor the owner-seller. This social loss can be avoided if the owner-seller can price discriminate between different users and charge different prices, according to their willingness and ability to pay. Under perfect price discrimination (every user pays according to his willingness to pay) the social losses would disappear but all revenue would go to the owner (the entire surface would become yellow). Conversely, in the absence of any property rights, all benefits may go to users; the "owner" would have no revenue or benefits (the blue triangle would cover the entire surface).

Figure 1a and 1b: Granting exclusive control rights on non-rival products



From a static welfare perspective that focuses on short-run price efficiency exclusive ownership rights are not a good idea. However, from a dynamic long-run perspective, this exclusive right creates an incentive for other agents to invest in innovation that can lead to new intellectual property rights. It generates a continuous stream of novel ideas and innovations that enhance social welfare for society as a whole. More exclusive rights are required when investments in innovation are costly. Exclusive ownership rights trade-off short-term price and welfare inefficiencies against long-term innovation and welfare gains.

How steep that trade-off should be is the subject of considerable debate in society because it confronts the interests of innovators with those of consumers and society at large. This is shown in Figure 1b. More protection always benefits the exclusive rights holder and increases his revenue. But too much protection inhibits spreading of the benefits across society and may generate welfare losses. For example, patents are limited in time in order to allow other inventors to build on previous inventions. They force the inventor to disclose his invention so that others can build on this innovation, though they cannot commercialise physical goods that make use of the patent without the authorisation of the rights holder.

In line with the institutional hypothesis on ownership rights, the optimal degree of protection may vary in function of changes in underlying prices. For example, when the innovation investment cost declines less protection is needed to achieve the same benefits and the optimal degree of protection. In the case of data, when the cost of investment in data collection declines with the introduction of digital technology, less protection is needed. In the extreme, when data is a pure by-product of on-going activities, no protection is required in order to incentivise data collection. However, this argument can also be turned around. When personal data is easy and cheap to collect, more exclusive personal data rights are required to protect the privacy of individuals.

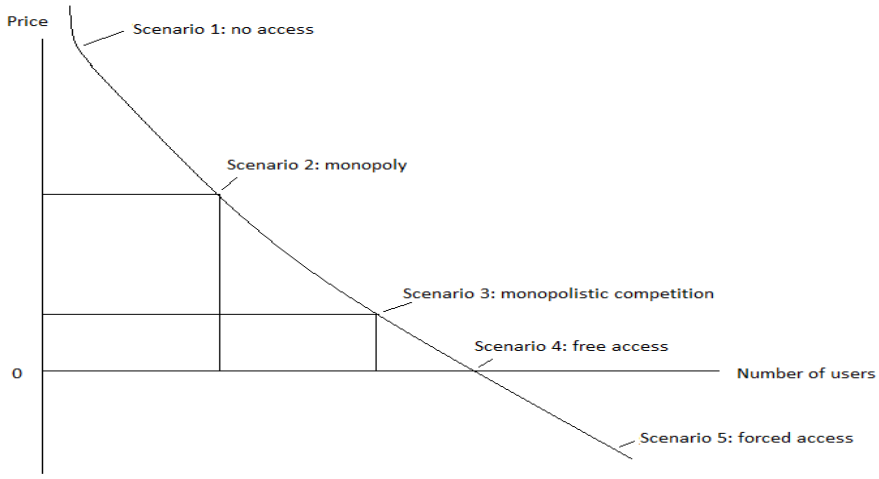
4. Modalities of data access and sharing

Trading data is inherently difficult: once they are revealed to a potential buyer to entice his interest the latter would no longer be willing to pay for the data because he already knows their information content. This is known as the Arrow Information Paradox (Arrow, 1962). This Paradox has triggered a vast volume of research on how to manage innovation by means of patents and copyright. Most of that is also applicable to access and trade in digital data and gave rise to the idea that data would also need some legal protection in order to facilitate access and trade. Without mentioning Arrow, this idea underpinned the European Commission's proposal for a data ownership right (European Commission, 2017). Granting a legally recognised ownership right would make it easier for the data holder to protect his rights *vis à vis* third parties and would therefore stimulate trade in data. Contemporary research on trading data in digital markets is essentially focused on finding market mechanisms that overcome Arrow's Paradox without granting legal *in rem* exclusive ownership rights (Bergemann and Bonatti, 2018). These mechanisms usually entail selling data or information in different qualities at different prices, and revealing only part of the available information in order to test the buyer's willingness to pay. The literature also distinguishes between direct data sales whereby the buyer receives a dataset and indirect data services sales whereby the buyer receives a service that is based on data – for example an advertising slot - but not the underlying dataset.

Below we explore a range of data access conditions, from fully closed to widely open, in function of the cost of access conditions for the user. In Figure 2 the horizontal axis represents the share of potentially interested users that get access to the data; the vertical axis represents the price at which access is granted. We assume here that the Arrow Paradox has been overcome by means of legal rights or bilateral contracts that are considered sufficiently robust to create a reliable setting for granting access. Access conditions create a cost for users and a benefit for the data holder. They trade off the interests of both parties. Data holders can be in a *de facto* exclusive data access situation. They can decide to sell the data at a monopolistic price, or keep them for their own use. In other situations there may be various substitute sources for data, making them less exclusive and that reduces the price of data in trading. Data holders may also chose to share data and make data

available for free[4]. In short, there is a wide variety in modalities and conditions for access to data, with potentially very different economic outcomes for individuals and for society as a whole. Regulatory intervention may try to push the market beyond voluntary exchanges towards further openness, or conversely restrict trade below the level of voluntary exchange. That may create costs for data holders and users. Society can choose to do so provided that the social benefits from access to the data outweigh the costs for individuals and firms.

Figure 2: Data access conditions



In Scenario 1 the data holder keeps the data for his own use and prevents any form of access, protected by legal and/or Technical Protection Measures. No user gets access and the price becomes infinitely high but the quantity sold is zero and revenue is zero. A firm may give temporary access to the data in order to allow others to experiment and develop innovative products and services with these data. When an innovative use becomes successful, the firm may decide to wrap this application back into its own fold and close off access to the data supply. This may be an ex-ante announced innovation strategy for large data holders (Parker, Van Alstyne, 2017).

Alternatively, rather than selling data directly, the data holder can keep them for his own exclusive use or for the production of a data-driven service that he sells to others. For example, Google does not directly sell or share consumer data but uses them for selling advertising slots on its Search engine and elsewhere. Consumers benefit from the information services while advertisers pay for the advertising service. In this way even non-accessible data can generate benefits for individuals and for society. That is reflected in Scenario 2 where the holder of exclusive data provides access at a revenue-maximizing monopoly price. Only some users get access to the data at this price. Users still earn a consumer surplus but there is a social deadweight loss from this monopoly. Society as a

[4] Note that the labels "data trade", "data sharing" and "access to data" are used interchangeable here: they all denote some form of data transfers between two parties.

whole loses because not all potential benefits from the data are realized. Welfare losses can be avoided when data monopolists can perfectly price discriminate between clients for the data (Bergemann & Bonatti, 2015). This requires information on the economic characteristic of the client. This situation may occur when the data user (the buyer) runs a downstream service that depends on a continuous supply of data from a monopolistic data holder. The data holder will be aware of the profit that the user makes on his downstream service and will try to appropriate that profit by increasing the price of the data and threatening to cut off the supply if the ask price is not paid.

The availability of alternative data sources reduces monopolistic rents (Scenario 3). When other sources are perfect substitutes, a competitive market emerges where the price will decrease to the marginal cost of producing and transferring the data. Since data production and collection is often subject to high fixed costs and almost zero marginal costs, this may push the price down to virtually zero. The marginal cost pricing rule is therefore unstable and not a good guidance in many digital economy settings. A more realistic description is monopolistic competition in data markets whereby several sources of data are partial or imperfect substitutes. For example, advertisers can find many sources of consumer profile information to target their online advertising. In this scenario, all those who are willing to pay the market price get access to the data. It may still leave some part of demand unsatisfied.

Data holders may of course decide to make the data freely available (Scenario 4) and voluntarily abandon their exclusive control over the data. They may see a benefit in doing so. For example, individuals may want to share some information on their private activities with a group of friends or even with the rest of the world through their social media pages and blogs because it benefits their social life. Firms may want to advertise their products and services on websites in order to increase sales. Data holders may make data freely accessible because it benefits them in other respects. Figure 2 shows how free access potentially benefits the widest possible group of users, similar to the perfect price discrimination case. The difference is that with perfect price discrimination all benefits are monetised by the data holder while with full free access benefits accrue to the downstream users. Both cases eliminate social welfare losses and generate the maximum benefit for society.

Data sharing may also come at a cost to the data holder when data are extracted against his will and at a cost to him (Scenario 5). Keeping some data private may have a social opportunity cost for society. How to distinguish between socially desirable and undesirable externalities? A simple criterion would be that, if the social value of data to society exceeds the private value of the data for individuals or firms, it would be beneficial for society to open access to the data for the public benefit. A straightforward example is the use of private data to detect or prevent crime or, more broadly, socially undesirable behaviour. Society may want to reduce data privacy rent-seeking behaviour when that inflicts a social cost. A controversial example is the use of face recognition by police in public spaces to deter citizens from deviant behaviour and improve public safety. Doing so may reduce socially undesirable behaviour and increase social welfare for all, but at the expense of privacy.

Less straightforward is the example of a search engine sending recommendations on a product of unknown quality in order to get feedback on its evaluation, or a navigation app sending drivers

deliberately into a road in order to collect information about the status of that road. The driver incurs a private cost when the road is not suitable for him but other drivers gain a social benefit from that information. Behavioural experiments with search engine and recommender systems happen every day on a large scale. They generate useful information for the system operator and possibly for the benefit of the community of users but may create costs for some users (Che & Horner, 2017).

The two main involuntary data access tools under EU law are the portability provisions under the GDPR and the text and data mining (TDM) exception to copyright. Portability permits data subjects to transfer their personal data from one service provider to another and thereby diminishes the provider's exclusive access to the data. It stimulates competition in downstream data-driven service markets. TDM on (publicly or privately) accessible data sources weakens monopolistic behaviour by data owners. In the EU, TDM cannot increase competition in data-based services however because commercial applications of the harvested data are not allowed. The US "fair use" and "transformative use" interpretations may be more flexible. US- based firms may thus have an advantage over EU-based firms in this respect.

Arrieta-Ibarra et al (2018) argue that there is underproduction of data and propose to give data producers an incentive to contribute to the production of data: Data as Labour. Data without remuneration offers no incentive to supply more data, other than data as by-products of on-going activities. That may be fine for B2C consumer-oriented data. For more productivity-oriented B2B data services users would have to go out of their way to produce more data. That requires remuneration as a production incentive. We already see this happening in many online labour platforms that are used to annotate digital data for use in AI/ML applications.

So far we have looked at simple data markets with transactions between two parties. We have not examined data trading from a platforms or multi-sided markets perspective. In multi-sided markets what happens on one side of the market affects the other sides. Data access can be free for one side but other sides may have to pay an access price. For example a widely applied data access model in B2C consumer markets is an exchange of data in return for watching ads. Google Search and Facebook give consumers free access to information services while advertisers pay for an advertising slot. They can target the ad to specific audiences but pay a price for this data-driven service without getting access to the underlying data. Clearly, data play a crucial role in delivering these services in multi-sided markets. However, the classic theory of multi-sided markets (see Martens 2017 for an overview) has no role for data in these markets. It relies only on network effects to work out differential access pricing strategies for each side of the market. Prüfer and Schottmueller (2017) present a model whereby platforms invest in data collection in order to get an informational advantage over their competitors. But so far we have no theory yet of the role that data and ML algorithms play in the operations of platforms.

5. How regulatory intervention affects access to data

Should we apply the same economic principles of IPR to data and assign exclusive access rights to data owners? How would that impact on the development of AI and ML? These questions have been at the centre of the data policy debate in the last couple of years. So far there are no exclusive ownership rights on data in the EU or elsewhere, though attempts have been made to develop IPR-like legal systems for data: the Database Directive.

The EU Database Directive (DBD) (1996) contains two provisions: a copyright and a sui generis right. Copyright protects the structure of databases[5] which, if original, constitutes the author's own intellectual creation. By contrast, the more controversial sui generis right protects databases regardless of their originality, as long as there has been "substantial investment in obtaining, verifying or presenting the contents"[6]. Both the copyright and the sui generis right in the DBD put the database owner in a monopoly position and therefore reduce access to the data. This database ownership right can be held up against third parties that have no contract with the data owner. It gives the database producer some security when data are in the public domain and avoids that they end up in commercial applications without any remuneration for the database producer. All ECJ court cases regarding the DBD were about the interpretation of the conditions that apply to invoke the ownership right. While the DBD grants exclusive rights it is also subject to exceptions that reduce the coverage of these rights. For example, the text & data mining (TDM) exception to copyright protection also applies to the DBD and constitutes a measure that opens access to databases. The EU TDM exception is limited to non-commercial use of the data. It is an opening combined with a restriction. In the US, the "fair use" provisions allow for wider use of data obtained through TDM, provided they do not compete with the services offered by the data holder.

An evaluation of the DBD (European Commission, 2018) claims that the sui generis right does not apply to databases that are by-products of the main activity of a firm, and therefore it does not apply broadly to the data economy (machine-generated data, Internet-of-Things devices, big data, AI, etc.). It concludes that "any meaningful move towards a policy intervention on the sui generis right would need to be substantial. It would need to take into account the policy debates around the data economy". In fact, the DBD's attempt to bring data ownership closer in line with the concepts that underpin intellectual property rights law has led to difficulties in the interpretation of the Directive (Hughenholtz, 2018). Data are rarely produced as a result of intellectual efforts; they are generated by observing human behaviour or states of nature, or by industrial processes steered by machines.

Probably the most important legal instrument in the EU is the General Data Protection Regulation (GDPR) that sets rules regarding the use of personal data. Any data that can be linked to a natural

[5] The DBD covers ownership rights to entire databases, not necessarily individual data points. For example, personal data included in a database are still subject to the provisions of the GDPR and therefore not the exclusive ownership of the database owner. A similar argument applies to copyright protected content included in a database.

[6] A recent evaluation of the DBD (2018) concludes that the DBD has no proven impact on the production of databases in the EU. However it has avoided legal fragmentation in the EU DSM and effectively harmonised EU legislation on databases (without any impact). Still it argues that the DBD provides an appropriate balance between protection of investments and interests of users.

person is considered to be personal data, irrespective of the source, storage or transmission mechanism. The GDPR gives data subjects a number of control rights over their personal data, including the right to obtain consent from the data subject before accessing the data, to right for data subjects to access their personal data and delete them, to port their data to other uses, etc. Taken together all these rights may come close to ownership rights but the GDPR does not give legal recognition to the concept of ownership rights on personal data. It prefers to give a number of specific rights rather than attribute full and tradable residual rights to an owner. Privacy is considered to be a basic and inalienable human right that cannot be traded away. For this reason, personal data are assumed not to be the subject of an economic transaction. That underlying principle has created tensions among policy makers, for example with regard to the legal interpretation of transactions whereby data subjects exchange their personal data in return for a service[7].

The GDPR grants data subjects' a number of rights over their personal data, including the right to withhold consent to access the data. It also prohibits use of personal data for other purposes than the originally intended purpose, unless they are anonymized. This makes it difficult for holders of personal data to aggregate them into larger databases or share them with other types of data services for the purpose of aggregation and analysis by means of ML algorithms. There are also provisions in the GDPR that work in the opposite direction and facilitate wider access to data. For example, Article 20 mandates portability of personal data. At the request of the data subject, the data service provider has to transfer the personal data to a destination of choice of the data subject. The data subject can thus chose to give wider access to his data. The GDPR also allows access to personal data without consent when required for technical reasons or because of legal obligations. An example is road traffic data. An important source of road traffic data is tracking of mobile phones in cars by telecom service operators. Tracking is a technical measure required to keep moving phones connected to the antennas of the mobile phone network. The GDPR allows this without consent because the data are necessary for the technical operations of the service. Moreover, the GDPR allows the onward transmission of the data provided they are anonymised, i.e. in the case of road traffic data aggregation they are stripped of the private phone number. Some authors suggest that tensions are emerging between the DBD and the GDPR regarding the definition of the borderline between personal and non-personal data (Graef et al, 2018).

A number of EU policy communications on data access are not legally binding. The European Commission Communication "Building the EU data economy" (January 2017) that introduced a new concept of "machine-generated data", created by computer processes or by sensors, and the introduction of an exclusive "data producer's right": a right to use and authorise the use of non-personal data for the owner or long-term user (i.e. the lessee) of a device. A more recent Communication "Towards a common EU data space" (April 2018) contains some normative

[7] Article 3 of the proposed EU Directive on the supply of digital content (COM(2015) 634 final) recognizes the contractual notion of digital content supplied not only for a monetary payment but also in exchange for (personal and other) data provided by consumers. However, the European Data Protection Supervisor pointed out that personal data cannot be the object of a trade contract (EDPS Opinion 4/2017).

"guidance" principles for B2B data sharing between private companies: transparency in contracting terms, recognition that several parties have contributed to creating the data and respect for the commercial interests of data holders and data users, ensuring undistorted competition in data markets, and facilitate data portability. It remains silent on the data ownership right proposal. It suggests that it can be socially efficient for companies to share data so that the value resulting from the data can be exploited to the maximum. The EU Public Sector Information (PSI) Directive seeks to make data held by public sector organisations, including commercial utilities for example, more accessible to citizens and firms. To the extent that some public sector organisations may be involved in the production of commercial services - for instance public utilities in transport and energy - it may also cover some commercial operations. The Directive suggests that pricing of data access should be at marginal costs. Digital datasets are often characterised by high fixed costs to create a dataset and virtually zero marginally costs to replicate and transmit it.

We conclude that regulatory intervention in EU data markets is currently rather limited. There are no exclusive ownership rights to data though there is a "database" ownership right with restrictive conditions. There are rights for data subjects regarding their personal data but they stop short of an exclusive ownership right. All these initiatives are bouncing back and forth between two poles: (a) offering more exclusive rights, either for the protection of personal rights or as an incentive to invest in data collection, and (b) making data more widely available and accessible to facilitate the extraction of new insights from data. This regulatory landscape leaves a wide open space where market-based data exchanges and transactions rule, based on bilateral contracts and TPMs for firms and individuals. In the next section we explore how that largely market-driven data space affects access to data, data aggregation and AI/ML.

6. Some tentative conclusions

The objective of this paper was to examine the interaction between digital data, data access regimes and the development of AI and its currently dominant technology, ML. In order to do so we started with the economic impact of digital data on decision making and explored how AI/ML could be fitted in that setting. Digitization triggered a steep drop in the cost of information. The resulting data glut created a bottleneck because human cognitive capacity is unable to cope with large amounts of information. Artificial intelligence and machine learning (AI/ML) triggered a similar drop in the cost of machine-based decision-making and helps in overcoming this bottleneck. Substantial change in the relative price of resources puts pressure on ownership and access rights to these resources. This explains pressure on access rights to data.

ML thrives on access to big and varied datasets. Access to data thus matters for the development of AI, at least in its current form of ML. The economic characteristics of data (non-rivalry, economies of scale and scope) favour data aggregation in big datasets. At the same time, non-rivalry implies the need for exclusive rights in order to incentivise data production when it is costly. The balance between access and exclusion is at the centre of the debate on data regimes. We explored the economic implications of several modalities for access to data, ranging from exclusive monopolistic

control to monopolistic competition and free access. We found that all these modalities allow for some form of data access though the cost of access and degree of sharing may vary substantially.

We then looked at how regulatory intervention may improve data access. It may push the market beyond voluntary exchanges, either towards more openness but possibly also towards further reduced access. This intervention may generate private costs for firms and individuals. Society can choose to do so if the social benefits of this intervention outweigh the private costs. The EU General Data Protection Regulation that defines the rights of data subjects with respect to their personal data and the Database Directive that grants exclusive ownership rights to database producers. Both instruments combine provisions that facilitate and reduce access to data. In this way they try to strike a balance between the private interests of data holders and the wider social interests of society. These two instruments leave a wide legal no-man's land where data access is ruled by bilateral contracts and Technical Protection Measures (TPMs) that give exclusive control to de facto data holders, and by market forces that drive access, trade and pricing of data. As such this data regime is incomplete because it does not grant exclusive and residual ownership rights for a large category of data, especially non-personal data. On the one hand the absence of exclusive rights might facilitate data sharing and access. That should benefit data aggregation and the production of big data sets that can be used for AI/ML applications. On the other hand, the absence of legal rights may result in de facto ownership enforced by TPMs and a segmented data landscape where aggregation into big datasets is hard to achieve. Data access rights matter for the downstream economic value generated by the use of data, in particular for value generated in data aggregation and use in ML applications.

We know that changes in access rights can have economic consequences for the welfare of private persons and firms as well as for overall societal welfare. It is unclear however what type of regulatory interventions would be required in a market-based data regime with incompletely specified ownership and access rights, in order to maximize the welfare of society and facilitate the development of AI/ML. Any intervention will have to balance several sides in the debate. This will require further research, in particular looking for empirical evidence on the static welfare and dynamic innovation impact of different data access regimes.

Bibliography

Agarwal Ajay, Joshua S. Gans and Avi Goldfarb (2018) Prediction, Judgment and Uncertainty, NBER working paper.

Agarwal Ajay, John McHale, Alex Oettl (2018) Finding needles in haystacks: AI and recombinant growth, NBER Working Paper 24541.

Athey, S. (2017). The impact of machine learning on economics. In Economics of Artificial Intelligence. University of Chicago Press. Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B., "Avoiding Discrimination through Causal Reasoning", NIPS 2017.

Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jimenez, Jaron Lanier and E Glen Weyl (2018) Should we treat data as labor? Moving beyond "free".

Bergemann, D. and A. Bonatti (2012), Markets for data, https://economicdynamics.org/meetpapers/2012/paper_538.pdf

Bergemann, D. and A. Bonatti (2015), Selling cookies, American Economic Journal: Microeconomics, 7(3), 1–37.

Bergemann, D., A. Bonatti and A. Smolin (2014), Selling experiments: Menu pricing of information, Cowles Foundation Discussion Paper No. 1952.

Brynjolfsson, Eric, Avi Ganamanemi and Felix Eggers (2017) Using Massive Online Choice Experiments to Measure Changes in Well-being, MIT digital economy initiative.

Cardona, M, N Duch-Brown, J Francois, B Martens, Fan Yang (2015) The macro-economic impact of e-commerce in the EU Digital Single Market, JRC Digital Economy working paper 2015-09.

Che Yeon-Koo and Johannes Horner (2017) Recommender Systems as Incentives for Social Learning.

Cockburn, Ian, Rebecca Henderson and Scott Stern (2018) The impact of AI on innovation, NBER Working Paper 24449.

Dengler, S and J Prüfer (2017) Consumers' Privacy Choices in the Era of Big Data, TILEC Discussion Paper No. 2018-014, Tilburg University.

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

European Commission (2017) Building the European Data Economy, Communication to the European Parliament and Council.

European Commission (2018) Towards a common European data space, Communication to the European Parliament and Council.

European Commission (2018) Artificial Intelligence: A European perspective, Joint Research Centre of the European Commission, December 2018, available at https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective

Graef, Inge, SihYuliana Wahyuningtyas and PeggyValcke (2015) Assessing data access issues in online platforms, Telecommunications Policy, vol 39.

Graef, Inge, Martin Husovec, Nadezhda Purtova (2018) Data Portability and Data Control: Lessons for an Emerging Concept in EU Law, mimeo Tilburg University.

Hugenholtz, Bernt (2018) Data property in the system of international property law: welcome guest or misfit? In Trading data in the digital economy: legal concepts and tools, Lohsse, Schulz and Staudenmayer, editors, Nomos Hart Publishing.

Kerber, Wolfgang (2016) A new intellectual property right for data: an economic analysis. GRUR 989.

Lee, Kai-Fu (2018) AI Superpowers.

Lewis R and J Rao (2015) The Unfavorable Economics of Measuring the Returns to Advertising, Quarterly journal of economics.

Martens, Bertin (2013) What does economic research tell us about cross-border e-commerce in the EU digital single market, JRC Digital Economy Working Paper 2013-05.

Martens, Bertin (2016) An economic policy perspective on online platforms, JRC Digital Economy working paper 2016-05.

Martens, Bertin and Frank Muller-Langer (2018) Access to car data and competition in aftersales services markets, JRC digital economy working paper (forthcoming)

North, Douglas (1994) Economic performance through time, American Economic Review.

OECD (2015) Data-driven innovation

Palfrey S and Urs Grasser (2012) Interop: the perils and promises of highly interconnected systems.
Parker, Jeffrey and Marshall Van Alstyne (2017) Six Challenges in Platform Licensing and Open Innovation.
Parker, Jeffrey and Marshall Van Alstyne (2017) Innovation, Openness, and Platform Control, Journal of Management Science.
Pilaszy I and D Tikk (2009) Recommending movies: even a few data are more valuable than metadata, Recsys, 2009.
Prufer, J and Christoph Schotmueller (2017) Competing with big data.
Rosen (1983) Specialization and Human Capital, Journal of Labour Economics.
Varian, H. R. (1997). Economic aspects of personal privacy. In Privacy and Self-regulation in the Information Age. US Department of Commerce
Zhu, H., S. E. Madnick and M. Siegel (2008), An economic analysis of policies for the protection and reuse of non-copyrightable database contents, Journal of Management Information Systems, 25(1), 199–232.

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

Joint Research Centre

EU Science Hub