11

Data Portability in a Data-Driven World

Frederike Zufall and Raphael Zingg

Today's technology giants have won market dominance through the collection, analysis, and synthesis of data. With the increasing dependence on digital technology, and increasing data dependency of said technology, data can be seen as a precondition to economic participation. Exploiting the steep economies of scale and network externalities of data, firms such as Google, Facebook, and Amazon are in positions of near monopoly. When such service providers disallow users from transferring their data to competing services, they can lock in users and markets, limiting the entry of market competition. Providing users with rights to both retrieve their data and transmit it to other firms potentially serves as a counterbalance, easing the acquisition of users for new market entrants. As such, data portability legislation has been claimed to have far-reaching implications for the private sector, reducing or hindering tools of forced tenancy. With users no longer married to a single firm, inroads for new technology are paved, with the average user more likely to have the ability and resource to change provider and adopt a solution that better suits their individual needs.

This chapter explores the concept of data portability in a world driven by artificial intelligence (AI). Section I maps out the journey that data takes in a data economy and investigates the valuation and cost of data. It posits that, because of data analytics and machine learning models, "generated" data, as data that has been derived or inferred from "raw" data, is of higher value in the data market, and carries a higher cost of production. Section II discusses what is required for the free flow of data in competitive datacentric markets: regulations on data tradability and portability. Our analysis leads to doubt that the newly introduced, hotly debated rules regarding portability of data under European Union (EU) law will adequately provide these prerequisites. The chapter concludes by suggesting an alternative model for data portability that distinguishes on a value basis rather than between personal and nonpersonal data.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

I THE JOURNEY AND VALUE OF DATA

This first section reviews the journey of data from collection to classification: the path from its moment of provision by a data subject to its subsequent transformation into inferred data. We present a categorization model that distinguishes data according to its origin, primarily distinguishing between raw and generated data. Utilizing these categories, we illustrate how data generated by machine learning models is being created at an exponential rate in today's data-driven economy. Lastly, a data valuation model is introduced, holding that the value of generated data is higher than raw data, and that the value of generated data scales exponentially in aggregation. We claim that the added value of generated data is created by firms that carry the costs of providing big data analytics, including machine learning.

A *Origin of Data*

Data can be classified according to a variety of parameters. A classification model can rely on the sensitivity of the subject, purpose of use, context of procession, degree of identifiability, or method of collection of data. We build on a categorization model of data that was introduced by a roundtable of Organisation for Economic Co-operation and Development (OECD) privacy experts in 2014,[1] and expanded by Malgieri.[2] The taxonomy categorizes data according to its origin – that is, the manner in which it originated – and distinguishes between raw data (provided and observed data) and generated data (derived and inferred data).

Raw data ("user-generated data") encompasses provided and observed data. Provided data is data originating from the direct actions of individuals (e.g., registration form filing, product purchases with credit card, social media post, etc.). Observed data is data recorded by the data controller (e.g., data from online cookies, geolocation data, or data collected by sensors).

Generated data ("data controller-generated data") consists of derived and inferred data. Derived data is data generated from other data, created in a "mechanical" manner using simple, non-probabilistic reasoning and basic mathematics for pattern recognition and classification creation (e.g., customer profitability as a ratio of visits and purchases, common attributes among profitable customers). Inferred data is data generated from other data either by using probabilistic statistical models for

[1] "Protecting Privacy in a Data-Driven Economy: Taking Stock of Current Thinking" (OECD, 21 March 2014), https://perma.cc/AFH5-MZF9 refers to provided, observed, derived, and inferred data – inferred data being defined as the "product of probability-based analytic processes".

[2] G Malgieri, "Property and (Intellectual) Ownership of Consumers' Information: A New Taxonomy for Personal Data" (2016) 4 *Privacy in Germany* 133; G Malgieri and G Comandé, "Sensitive-by-Distance: Quasi-Health Data in the Algorithmic Era" (2017) 26 *Information & Communications Technology Law* 229; Boston Consulting Group, "The Value of Our Digital Identity" (2012), distinguishes between volunteered, required, tracked, and mined data; World Economic Forum, "Personal Data: The Emergency of a New Asset Class" (2011), distinguishes between volunteered, observed, and inferred data.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

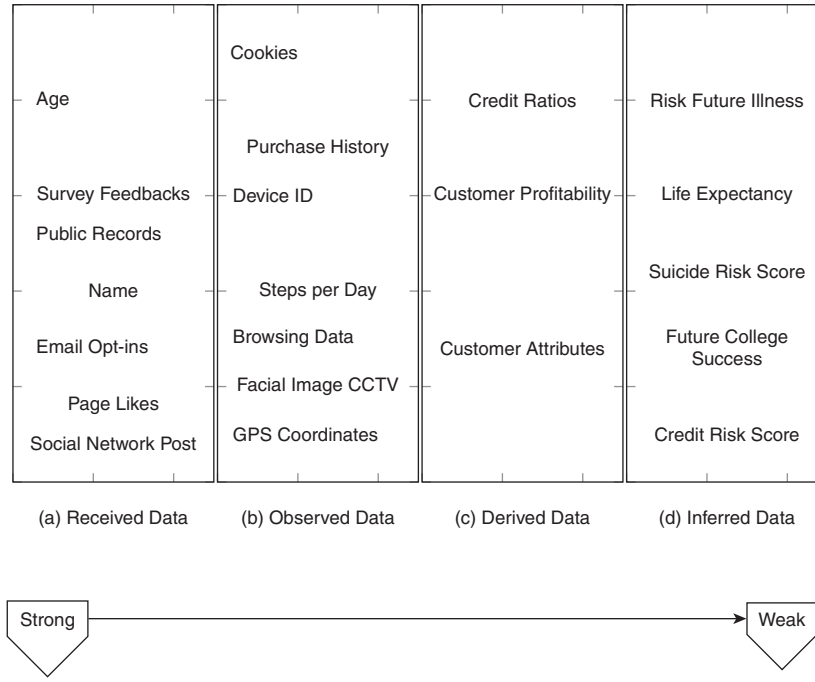

FIGURE 11.1 Data types and examples

testing causal explanation ("causal inferences") or by using machine learning models for predicting output values for new observations given their input values ("predictive inferences").[3]

The relationship between information and the data subject can be classified as either strong (provided data), intermediate (observed and derived data), or weak (inferred data). The stronger the relationship, the more individuals are involved in the creation of the data. Illustratively, a Facebook user has a strong relationship with their registration data and their posts. An example of a weaker relationship would exist when Facebook, based on its algorithmic models, assigns a liberal or conservative political score to this user. The user's age, geographic location, and posts are all data provided by the user, and eventually included as independent variables in the model. But it is Facebook's model that will ultimately predict the likelihood the user belongs to either group.

The evolving relationship from provided to inferred data, or from a weak to strong relationship between the data subject and the data, is illustrated in Figure 11.1. Although the delimitation of data types is crucial, a number of gray areas exist. Take the example of a data subject that does not upload data themself, but actively selects which sets of data and their conditions the data controller may access. It is unclear

[3] G Shmueli, "To Explain or to Predict" (2010) 25(3) *Statistical Science* 289.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

whether these datasets are received or observed by the controller.[4] Note that inferred data is created not only by the analysis of a specific user's data, but also by the analysis – via statistical learning and automatic techniques to elicit patterns – of all data available to the data generator, including personal data provided by other users.[5]

B *Artificial Intelligence and Data*

With the rise of AI, generated data is expected to proliferate at an exponential rate. As more and more institutions take advantage of increasingly broad datasets, computing power, and mathematical processes,[6] the amount of generated data will expand and the costs of prediction decrease.[7] As pointed out by a recent report ordered by the House of Commons of the United Kingdom, protecting data helps to secure the past, but protecting inferences is what will be needed to protect the future.[8] Inferential data generated by machine learning techniques has already been used (with varying success)[9] to predict sexual orientation based upon facial recognition; emotions of individuals based on voice, text, images, and video; a neighborhood's political leanings by its cars; and physical and mental predictions, to name but a few.[10] With the advancement of the technology and the availability of large training sets, the accuracy of inferred predictions will increase as well.

The predictive potential of machine learning is not confined to academic use cases, as commercial applications abound. Recent patent applications in the USA include methods for predicting personality types from social media messages,[11] predicting user behavior based on location data,[12] predicting user interests based on image or video metadata,[13] or inferring the user's sleep schedule based on smartphone and communication data.[14] In all these instances, raw user data is collected on mobile devices

[4] G Malgieri, "'User-Provided Personal Content' in the EU: Digital Currency Between Data Protection and Intellectual Property" (2018) 32(1) *International Review of Law, Computers & Technology* 118.

[5] R Accorsi and G Müller, "Preventive Inference Control in Data-centric Business Models" (2013), https://perma.cc/T722-JM47.

[6] "Analytics Comes of Age" (2018), https://perma.cc/MV8Y-3M5B; M Abrahams, "The Origins of Personal Data and Its Implications for Governance" (2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2510927.

[7] J O'Callaghan, "Inferential Privacy and Artificial Intelligence: A New Frontier?" (2018) 11(2) *Journal of Law & Economic Regulation* 72.

[8] "Disinformation and 'Fake News': Final Report" (2018) Eighth Report of Session 2017–19, HC 1791, https://perma.cc/ZPK4-WB9J.

[9] O Etzioni, "No, the Experts Don't Think Superintelligent AI Is a Threat to Humanity" (*MIT Technology Review*, 20 September 2016), https://perma.cc/B543-HZZ5.

[10] See O'Callaghan, note 7 above.

[11] US10013659, "Methods and Systems for Creating a Classifier Capable of Predicting Personality Type of Users", granted 3 July 2018.

[12] US20170255868, "Systems and Methods for Predicting User Behavior Based on Location Data", filed 3 March 2017.

[13] US9798980, "Method for Inferring Latent User Interests Based on Image Metadata", granted 24 October 2017.

[14] US20160292584, "Inferring User Sleep Patterns", filed 31 March 2015.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

(e.g., smartphones, tablets, etc.) to build and train the predictive model, and then used to predict individual user characteristics and behaviors (as generated data). This generated data is of value to marketing and advertising firms or organizations more generally to identify target users for their products and services.

C *Valuation of Data*

In general, the valuation of data is difficult, as it varies widely by type, scale, and industry sector. We make two assumptions that underly this chapter, and that support our position that generated data is of higher value than raw data. We claim that the higher value of generated data derives from the investment of firms in development, and subsequent use of statistical and machine learning models.

Our first assumption is that at the single datapoint level, raw data is on average of lower value than generated data. Our explanation for this assumption is as follows: raw data (such as the age of a data subject) is assumed to be, on average, of lower value to companies than generated data (such as future health predictions). In fact, in the marketplace, the price for general information, such as age, gender, and location, can be purchased for as little as $0.0005 or $0.50 per 1,000 people.[15] We assume that the price for creation of and access to generated data is higher.[16] The value of the datapoint integrates the value-added created by the respective algorithm. This is a generalizable claim despite specific and highly contextual differences. To provide a counterexample, data relating to diseases directly provided by a patient might be of higher value to an insurance company than a prediction based on that data.[17]

Our second assumption is that, on a large scale, the value of raw data increases linearly, whereas the value of generated data increases exponentially. We make this assumption for the following reasons: for statistical or machine learning approaches, received and observed data will need to be purchased on a large scale in order to

[15] E Steel et al., "How Much Is Your Personal Data Worth?" (*Financial Times*, 12 June 2013), https://perma.cc/EDY4-7G6U. On the cryptocurrency marketplace, where the user can monetize their data directly, selling GPS location data (to Datum), Apple Health data (to Doc.ai), and biographical Facebook information and Strava running routes (to Wibson) will yield an estimated $0.3; see G Barber, "I Sold My Data for Crypto. Here's How Much I Made" (*WIRED*, 17 December 2018), https://perma.cc/AV2V-B9G2.

[16] B Ehrenberg, "How Much Is Your Personal Data Worth?" (*The Guardian*, 22 April 2014), https://perma.cc/MS7V-5R3W ("[t]he inferred data is the type with real practical value, and the first two, unsurprisingly, don't cost much; they just help to build a picture of the third"). D Ciuriak, "Unpacking the Valuation of Data in the Data-Driven Economy" (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3379133 (contending that "[t]he major share of the market value of these firms is comprised of IP [intellectual property] and data – arguably, mostly data, although there is no empirical basis for venturing a specific point estimate").

[17] With a price point at $0.3 per name for a list with names of individuals suffering from a particular disease, see P Glikman and N Glady, "What's the Value of Your Data?" (*TechCrunch*, 14 October 2015), https://perma.cc/M3GB-BA78.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

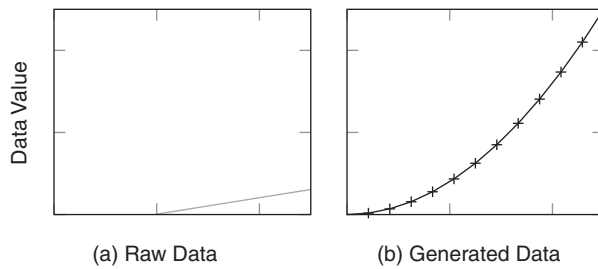

FIGURE 11.2 Data value of raw data and generated data

build models that will create inferred data. Since the accuracy of predictions is largely a function of the size of training datasets, we can assume that the value of received and observed data is close to zero for small-scale datasets. On the other hand, past acquisitions of datacentric companies reveal a significantly higher value per user, varying between $15 and $40. For instance, Facebook acquired WhatsApp and Instagram for $30 per user.[18] These per-user valuations reflect both the quality and scope of the information collected, as well as the expectation of continued platform engagement, and subsequent additional data creation by each acquired user.[19] In short, these acquisitions aim at exploiting trends and patterns in large groups with high confidence in the quality of the data. The process of value creation directly depends on investment in machine learning models needed to convert data into predictions.[20] These include direct operating costs, such as the employment costs of engineers, the licensing costs for software programs, the costs for obtaining more computer power and storage, and the costs of integrating systems with the implementation platform, as well as indirect costs such as training and change management costs or cybersecurity monitoring costs.[21] Therefore, the valuation of datacentric companies reflects the value of aggregated generated data or the potential for firms to create aggregated generated data.[22] We represent the respective value of raw data and generated data in Figure 11.2: with more data, the value of raw data increases linearly (a), whereas the value of generated data increases exponentially (b).

[18] G Sterling, "What's the Most Expensive (Per User) Acquisition? Hint: Not WhatsApp" (Marketing Land, 26 February 2014), https://perma.cc/KR3D-DWMQ.

[19] Glikman and Glady, note 17 above.

[20] DQ Chen et al., "How the Use of Big Data Analytics Affects Value Creation in Supply Chain Management" (2015) 32 *Journal of Management Information Systems* 4 (showing that the use of big data analytics explained 8.5 percent of the variance in asset productivity and 9.2 percent of the variance in business growth).

[21] N Cicchitto, "What Is the Cost of Implementing AI Today?" (Avatier, 9 May 2019), https://perma.cc/G8ML-WF8Q.

[22] See D Ciuriak, note 16 above (contending that the only comprehensive way to data valuation is to infer it from the market capitalization of data-driven firms).

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

II PORTABILITY OF DATA

This second section analyzes the newly introduced regulation of data portability in Europe. With the goal of moving toward a single market for data, the EU has sought to remove obstacles to the free movement of data via two regulations regarding personal and non-personal data. We evaluate the newly introduced right to data portability under the General Data Protection Regulation (GDPR)[23] and the porting of data regime under the Non-Personal Data Regulation (NPDR).[24] Our analysis of both data portability concepts suggests that the current separation between personal and non-personal data does not provide for a comprehensive and coherent data portability regime.

A *Free Flow of Data*

EU law has a long tradition of shaping regulation to create a single market for goods, services, people, and capital. In recent years, the European Commission has emphasized the need for a data ecosystem built on trust, data availability, and infrastructure.[25] Ensuring the free flow of data is part of this effort to establish a "digital single market".[26] Data is increasingly seen as a tradable commodity.[27] While the framework for trading data can be found in the traditional civil law rules for purchase contracts, the contract performance – that is, the actual transfer of the data – largely depends on the existence of data portability as a legal institution.[28] We are interested in how a regulatory framework for the market may level the playing field, challenging large incumbents with a vested interest in not transferring potentially valuable data to competitors.[29]

The more data is concentrated in the hands of a provider, the more likely it will be considered to hold a dominant position under EU competition law.[30] Although the

[23] 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 (hereinafter GDPR), OJ L 119, 4.5.2016, 1–88.

[24] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (hereinafter NPDR), OJ L 303, 28.11.2018, 59–68.

[25] "The Single Market in a changing world: A unique asset in need of renewed political commitment", COM/2018/772 final.

[26] F Zufall, "Digitalisation as a Catalyst for Legal Harmonisation: The EU Digital Single Market" (2017) 10 *WIAS Research Bulletin* 103.

[27] D Ciuriak, note 16 above.

[28] There is an ongoing debate surrounding the tradability of digital goods, and whether a resale is admissible or not; see H Zech, "Data as a Tradeable Commodity", in A De Franceschi (ed.), *European Contract Law and the Digital Single Market: The Implications of the Digital Revolution* (Cambridge, Intersentia, 2017).

[29] M Peritz and H Schweitzer, "Ein Neuer Europäischer Ordnungsrahmen für Datenmärkte?" (2018) 71 *Neue Juristische Wochenschrift* 275, at 277–278.

[30] Treaty on the Functioning of the European Union (TFEU), Article 102. Traditionally, the European Court of Justice considered the threshold to be at 40 percent or more market share; see Judgment of 13 February 1979 (*Hoffmann-La Roche & Co. AG v. Commission of the European Communities*), Case 85/76, ECLI:EU:C:1979:36.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

dominant competition law test is based on market share, not data concentration, said concentration is likely to lead to large market shares in data-driven markets.[31] The European Data Protection Supervisor has discussed how portability of data can foster a functioning market by preventing the abuse of dominance and the lock-in of consumers.[32] EU competition law, however, can generally be characterized as an *ex post* regulation: in fact, the European Commission only intervenes once a dominant position has been abused in already existing markets.[33]

As digital markets are especially prone to winner-takes-all (or -most) outcomes,[34] additional *ex ante* regulations are key. The EU has set up a number of these *ex ante* mechanisms, in particular in sector-specific regulation. A prominent example of this is the telecommunications sector: the Universal Service Directive established a right to number portability, considered a predecessor to the right to data portability under EU law.[35] The portability of telephone numbers and of data facilitates effective competition and can be considered a form of *ex ante* regulation as it creates the prerequisites for establishing a functioning telecommunication market.

The free movement of data is further addressed in Art. 16(2)1 of the Treaty on the Functioning of the European Union (TFEU), which gives the EU legislator the power to establish rules regarding the protection and free movement of personal data. The GDPR confirms the free movement of data as a subject-matter of the regulation and postulates that the free movement of personal data within the EU shall not be restricted or prohibited for the protection of personal data.[36] These affirmations refer once more to the foundation of the EU: free movement of goods, services, people, capital, and now data in a single market. Since May 2019, the regime is complemented by the NPDR.[37] Targeting non-personal data, the NPDR is entirely based on the ideal of the free flow of data. According to the NPDR, the two regulations provide a coherent set of rules that cater for free movement of different types of data.[38]

[31] See the recent decision of the German Federal Court of Justice: BGH, 23.06.2020 (KVR 69/19) confirming the assessment of the German Federal Cartel Authority, BKartAmt, 06.02.2019 (B6-22/16).

[32] "Preliminary Opinion of the European Data Protection Supervisor, Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy", March 2014 (hereinafter EDPO, Opinion 2014), paragraph 82–83 with reference to K Coates, *Competition Law and Regulation of Technology Markets* (Oxford, Oxford University Press, 2011).

[33] Although competition law, on the flipside, creates incentives for firms to adjust their behavior under threat of enforcement.

[34] E Brynjolfsson and A McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (New York, W.W. Norton & Company, 2014).

[35] Universal Service Directive (2002/22/EC), Article 30. EDPO, Opinion 2014, note 32 above, paragraph 83.

[36] GDPR, Article 1(1) and 1(3).

[37] NPDR, note 24 above.

[38] NPDR, Recital (10).

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

B *Regimes for Data Portability*

The portability of data is explicitly covered by both the GDPR and the NPDR. The former only applies to "personal data", the latter to "non-personal data".[39] EU law therefore clearly delineates personal from non-personal data.

Personal data is defined as "any information relating to an identified or identifiable natural person ('data subject')".[40] The notion of personal data is broad, as it only requires that a natural person can be identified directly or indirectly. It is sufficient, for instance, that the link to the natural person can be established using other reasonably accessible information – such as a combination of specific browser settings used to track behavior for personalized advertising.[41]

Non-personal data, by contrast, is any information that does not relate to an identified or identifiable natural person. Firstly, this encompasses data that originally does not relate to an identified or identifiable natural person, such as weather information or data relating to the operation of machines. Secondly, properly anonymized data cannot be attributed to a specific person and is therefore non-personal.[42] However, if non-personal data can be linked to an individual, the data must be considered personal.[43]

1 Portability of Personal Data

The newly introduced right to data portability in Art. 20 GDPR gives the data subject the "right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided". Data subjects shall have the right to receive the personal data concerning them and transmit that data to other controllers.

The provision mirrors the GDPR's dual purpose: both the protection of personal data and the free flow of personal data. The right to the protection of personal data is intertwined with a market-centered economic approach to personal data.[44]

[39] GDPR, Article 2(1).

[40] GDPR, Article 4(1).

[41] See for IP addresses: ECJ, Judgment of 24.11.2011 – C-70/10 – Scarlet/SABAM; ECJ, Judgment of 19.10.2016 – C-582/14 – Breyer/BRD.

[42] "Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union, May 29, 2019", COM (2019) 250 final (hereinafter Guidance on NPDR). See further subsection C.

[43] See M Finck and F Pallas, "They Who Must Not Be Identified: Distinguishing Personal from Non-Personal Data under the GDPR" (2020) 10 *International Data Privacy Law* 11; and I Graef et al., "Towards a Holistic Regulatory Approach for the European Data Economy: Why the Illusive Notion of Non-Personal Data Is Counterproductive to Data Innovation" (2019) 44 *European Law Review* 605.

[44] GDPR, Article 1(1); see P De Hert et al., "The Right to Data Portability in the GDPR: Towards User-centric Interoperability of Digital Services" (2018) 34 *Computer Law & Security Review* 193;

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

Not all personal data is subject to the right of portability. Only personal data for which the processing is based on consent or a contractual relationship is covered by the norm.[45] This limitation largely corresponds to the requirement that the personal data in question was provided by the data subject.[46] Accordingly, raw personal data is covered by portability because provided data is, by definition, directly provided by the data subject and observed data is (by most accounts) considered as such.[47] Generated data, however, whether derived or inferred, is not considered as being provided by the data subject.[48] Therefore, a large share of personal data is not subject to portability as it is not provided by the data subject.[49]

The GDPR provides a relatively strong right to data portability for the data subject. Data portability is seen from the data subject's perspective, with a focus on data protection. Creating a comprehensive regime for the portability of all kinds of personal data was not the priority of the EU legislator, as shown by the exclusion of generated personal data. Although the norm is often discussed as being situated in the area of competition law – with its aim of facilitating the free flow of data – data portability under the GDPR is still being considered closer to genuine data protection law than to regulatory competition law.[50]

2 Portability of Non-personal Data

With the NPDR, the EU encourages the porting of non-personal data.[51] The Internet of Things or industrial settings are major sources of non-personal data, as exemplified by aggregate and anonymized datasets used for big data analytics, data on precision farming, or data on maintenance needs for industrial machines.[52] The Regulation addresses two obstacles to non-personal data mobility: data localization

T Jülicher et al., "Das Recht auf Datenübertragbarkeit – Ein datenschutzrechtliches Novum" (2016) *Zeitschrift für Datenschutz* 358.

45 Typical cases are the creation of a social media profile based on personal data or providing shipping and billing information to an online shop (GDPR, Art. 6(1)(a) and (b)). Besides the data subject's right to data portability, the GDPR also takes into consideration the rights and protection of third parties as a limiting factor in case the datasets contain their personal data (GDPR, Art. 20(4)).

46 As the GDPR uses different phrasing for "personal data" and for "data provided by the data subject", the latter must be a part of the former.

47 "Guidelines on the right to data portability 16/EN/WP242rev.01" (hereinafter Guidelines on DP).

48 Ibid.

49 S Wachter and B Mittelstadt, "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI" (2018) 2 *Columbia Business Law Review* 494. For the practical challenge of porting Facebook data, see G Nicholas and M Weinberg, "Data Portability and Platform Competition: Is User Data Exported from Facebook Actually Useful to Competitors?" (2019), https://perma.cc/54RV-3G6G.

50 In this sense, Guidelines on DP, note 47 above, at 4. See further EDPO, Opinion 2014, note 32 above, paras 26, 83; W Kerber, "Digital Markets, Data, and Privacy: Competition Law, Consumer Law and Data Protection" (2016) *GRUR International* 639; Jülicher/Röttgen/v. Schönfeld, note 44 above; Peritz and Schweitzer, note 29 above, at 275, 277, 278.

51 NPDR, Article 6.

52 NPDR, Recital (9).

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

requirements imposed by the public sector and private vendor lock-in practices.[53] Such a lock-in effect might exist if cloud services like data storage or cloud-based data applications do not ensure the portability of the respective data.

While the GDPR provides for an enforceable right of the data subject, the NPDR approaches portability differently. The regulation encourages self-regulatory codes of conducts; that is, legally nonbinding instruments. The norm expressively refers to best practices that should facilitate the porting of data through "structured, commonly used and machine-readable formats including open standard formats where required or requested by the service provider receiving the data".[54] Meanwhile, codes of conduct on the porting of data and switching between cloud service providers have been developed by the cloud switching and porting data working group (SWIPO) for Infrastructure-as-a-Service (IaaS) and Software-as-a-Service (SaaS) cloud services.[55] These codes require, inter alia, the use of application programming interfaces (APIs),[56] open standards, and open protocols by cloud service providers.

C *Analysis: The Concept of Data Portability*

Our analysis depicts the limitations of existing EU law in providing for the free movement of data via a comprehensive and effective portability regime. In particular, we discuss how the distinction between personal versus non-personal data and raw versus generated data may impact the concept of data portability.

1 Distinction between Personal and Non-personal Data

The EU framework separates data into two types: personal and non-personal. This separation subjects data to different regulatory regimes – with a number of consequences in terms of portability. The distinction between personal and non-personal data is meant to preserve a high level of protection for data that can be related to an individual. The GDPR accordingly sets forth a right to access available for all type of personal data, whether raw or generated.[57] The NPDR targets data that is not related to an identifiable natural person. The interests of datacentric businesses stand in the center of the regulation. The free flow of data is therefore targeted from the data subject's perspective as well as from the perspective of market regulation.

[53] Compare NPDR Article 4 and Article 6.

[54] NPDR, Article 6(1)(a).

[55] See "Presentation of Codes of Conduct on Cloud Switching and Data Portability" (European Commission, 9 December 2019), https://perma.cc/H46G-33WN. Platform-as-a-Service might be considered at a later stage: "Cloud Stakeholder Working Groups Start Their Work on Cloud Switching and Cloud Security Certification" (European Commission, 16 April 2018), https://perma.cc/K2TT-DJKN.

[56] See, on the role of APIs, O Borgogno and G Colangelo, "Data Sharing and Interoperability Through APIs: Insights from European Regulatory Strategy" (2018) European Union Law Working Paper No. 38, https://ssrn.com/abstract=3288460.

[57] GDPR, Article 15.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

In theory, the distinction between personal and non-personal data appears straightforward. In practice, this is often not the case. For instance, large datasets where personal and non-personal data are mixed up (so-called mixed datasets) make it hard to identify the applicable legal regime. The NPDR recognizes this situation and addresses it by splitting up the application of both legal regimes to the respective type of data.[58] In cases where both types are inextricably linked, the application of the GDPR takes precedence (even if personal data represents a small part of the set only).[59] Addressing the complexity of GDPR compliance for mixed datasets can have a large impact on technology firms' associated costs. Uncertainty still prevails in the field on how to avoid falling under the GDPR.

This lack of legal certainty provides an incentive to anonymize data. The underlying belief is that personal data can be turned into non-personal data by anonymization, as anonymization destroys the link to an identifiable person. Consequently, the NPDR takes into consideration future technological developments making it possible to turn anonymized data into personal data, with the consequence of then having to treat such data as personal data and to apply the GDPR to it.[60] Recent studies, however, have challenged the common understanding of anonymization. The European Commission itself has addressed these concerns, but remains committed to the belief that anonymization can be achieved in practice.[61] In an influential study, Rocher, Hendrickx, and de Montjoye showed that 99.98 percent of Americans would be correctly reidentified in any dataset using fifteen demographic attributes.[62] A range of additional studies have supported this point, with reidentification of supposedly anonymous datasets in healthcare, ride-sharing, subway, mobile phone, and credit card datasets.[63] All this raises doubts about whether the distinction between personal and non-personal data can be upheld in the future.

2 Distinction between Raw and Generated Data

The right to data portability under the GDPR only applies to personal data provided by the data subject. From the viewpoint of providing access to the market of social media services, the portability of raw data alone is considered sufficient to prevent customer lock-in. Although the controller uses raw data (provided and observed) to generate derived and inferred data, generated data is not considered as "provided by the data subject" in the sense of Art. 20(1) GDPR. As such, generated data does not fall under the right to data portability. However, if it qualifies as personal data,

[58] NPDR, Article 2(2).
[59] Guidance on NPDR, note 42 above, at 8–10.
[60] NPDR, Recital (9).
[61] Guidance on NPDR, note 42 above.
[62] L Rocher et al., "Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models" (2019) 10 *Nature Communications* 1; see also C Blackman and S Forge, "Data Flows: Future Scenarios" (2017), https://perma.cc/QN7C-YRPZ at 22, box 2.
[63] For a summary, see Rocher et al., note 62 above; see further Finck and Pallas, note 43 above.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

generated data is still subject to the right of access or the right to not be subject to automated individual decision-making.[64] Consequently, the GDPR offers the data subject access to its personal data and protection regardless of whether the data is raw or generated.[65]

A reason why the right to data portability under the GDPR does not cover data created by the controller (i.e., generated data) might be that portability would here grant a strong advantage to competitors. Porting generated data would grant companies access to especially valuable data (Assumption 1), whose aggregation scales exponentially (Assumption 2).[66] The GDPR envisages a model where the data subject provides raw data to social media providers and leaves the additional value of the data to these providers as a compensation of their costs. But this is only justified in instances like Facebook, where the user "pays" with their data in exchange for the free use of the service. The service provider bears the cost of providing the social network and may recoup their investments by gaining profit from the added value the raw data gains through inferential information, illustratively via advertising. If the data subject, however, pays for a service, be it social networking or an analysis of their personal data, the situation is entirely different: the service provider's costs are being compensated by monetary payment. The added value of the derived or inferred data should then remain with the data subject and fall under the scope of the right to data portability.[67]

This situation is similar to the one envisaged by the NPDR: one between a customer and a service provider. When the customer provides raw data to the data service provider who conducts statistical analysis or prediction through machine learning on this data on behalf of the customer, the customer bears the cost of transformation of the data. As the costs are assigned to them, they should be able to obtain the value of the resultant generated data, to transfer it and switch providers. This right would in general already be subject to a civil law contract by which the relationship between service provider and customer is governed. The role and task of regulation would then only be to enforce portability in cases where service providers have market power to the extent that such portability and its conditions (portable file format, interfaces, etc.) is not subject to the service agreement. For this reason, the data porting rules under the NPDR may be insufficient as they are nonbinding and limited to self-regulatory measures. The European Commission or the respective member state authorities would need to take competition law measures based on abuse of dominant position, which have the limitation of being *ex post* in nature.

64 GDPR, Articles 15 and 22.

65 Compare the Opinion of the Committee on Legal Affairs for the Committee on Civil Liberties, Justice and Home Affairs, A7-0402/2013, PE501.927v05-00, at 520 ("This new right [to data portability] included in the proposal for a directive brings no added value to citizens concerning right of access").

66 See Section I, subsection C.

67 All the more if the service provider in these cases might not be a controller in the sense of Art. 4 (7) GDPR anymore, if the decision-making power is assigned to the data subject.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

An already binding obligation of non-personal data portability can be seen in Art. 16(4) Digital Content Directive,[68] albeit limited to the area of digital content and digital services: the consumer is granted the right to request "any content other than personal data, which was provided or created by the consumer when using the digital content or digital service supplied by the trader". This stipulation affirms our position: the value of digital content – that is, the data – created by the customer is assigned to the customer, leading to a right to retrieve that data in a commonly used and machine-readable format, as the second subparagraph states.

D *Data Portability beyond the European Union*

The EU has taken the lead in shaping the way the world thinks about data protection, privacy, and other areas of digital market regulation.[69] Its data protection standards in particular have been diffused globally.[70] Firstly, the ideas and concepts of the GDPR – in our case of data portability – have influenced a number of jurisdictions to enact data portability norms themselves. Secondly, international firms are bound directly by the GDPR's and the NPDR's extraterritorial scope, even without being established in the EU. Thirdly, because of the "Brussels Effect" foreign corporations often prefer to respect EU law even without a legal obligation to do so. Fourthly, international soft law has been and can be deployed to integrate data privacy principles from the EU, playing thereby a key role in the international governance of portability regimes. Fifthly, data privacy obligations have been stipulated in international treaties, requiring the implementation of data portability norms within the national law of ratifying states. In this regard, international economic law can help to diffuse data portability rules across the world.

1 Adoption by Third Countries

Numerous data protection laws around the world have emulated the GDPR, including its right to data portability.[71] A prominent example is the California Consumer Privacy Act, signed a month after the GDPR came into effect. The legislation incorporates portability in the context of the right to access as

[68] Directive (EU) 2019/770 of the European Parliament and of the Council of 20 May 2019 on certain aspects concerning contracts for the supply of digital content and digital services, OJ L 136, 22.5.2019, 1–27.

[69] See, on the idea of a "digital single market", "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Digital Single Market Strategy for Europe", COM (2015) 192 final; and Zufall, note 26 above, at 103–110.

[70] A Bradford, *The Brussels Effect: How the EU Rules the World* (New York, Oxford University Press, 2020).

[71] G Greenleaf, "The Influence of European Data Privacy Standards Outside Europe: Implications for Globalization of Convention 108" (2012) 2 *International Data Privacy Law* 68, at 77.

a modality of how electronic access should be provided; that is, in a portable format.[72] In comparison to the GDPR, the stipulation has an arguably broader scope, as all personal data is portable, and not only the personal data provided by the data subject. On the other hand, the norm is weaker as the businesses collecting personal information can provide access nonelectronically by simple mail, even if the data is stored digitally. Businesses are thus offered a way to circumvent portability, unless they do not mind the additional costs of mail delivery (which might be less than investing in interoperability).

Other examples of adoption include Benin, which enacted a GDPR-like legislation with its *Code du numérique* and included a right to data portability.[73] Brazil has adopted a new General Data Protection Law that introduces a right to the portability of data.[74] A possible codification of data portability is further vividly discussed by a number of countries.[75] Japan, for instance, has initiated a study to assess the merits and demerits of data portability, taking into consideration the costs for firms to establish portability.[76]

2 Extraterritorial Application

EU law imposes itself on foreign entities by extending its scope of application beyond EU territory. Inspired by the famous Google Spain judgment of the European Court of Justice,[77] Art. 3(2) GDPR introduces a remarkably broad territorial scope: GDPR applies to controllers or processors not established in the EU if the processing activities are related to the offering of goods or services to data subjects in the EU or to the monitoring of their behavior within the EU.[78] Data portability can therefore be requested by an EU citizen or resident from a foreign – for instance, US – firm, if the activities of the firm fall under the GDPR.

[72] The California Consumer Privacy Act (CCPA) of 2018 (Assembly Bill No. 375), Division 3, Part 4, Section 1798.100 (c) of the Civil Code ("The information may be delivered by mail or electronically, and if provided electronically, the information shall be in a portable and, to the extent technically feasible, in a readily useable format that allows the consumer to transmit this information to another entity without hindrance").

[73] G Greenleaf, "Global Data Privacy Laws 2019: 132 National Laws and Many Bills" (2019) 157 *Privacy Laws & Business International Report* 14.

[74] Federal Law No. 13,709 of 14 August 2018 (General Law for the Protection of Personal Data).

[75] "OECD Expert Workshop on Enhanced Access to Data: Reconciling Risks and Benefits of Data Re-Use" (2018), https://perma.cc/R673-W629; "Key Issues for Digital Transformation in the G-20" (2017), https://perma.cc/WJK5-PVU8.

[76] https://perma.cc/NG7U-K649; Japan's Ministry of Economy, Trade and Industry (METI) (2017), "Future Vision towards 2030s", full text in Japanese: https://perma.cc/AQK8-JLP2, at 204.

[77] Judgment of the Court (Grand Chamber), 13 May 2014, Case C-131/12 – *Google Spain SL, Google Inc v. AEPD, Mario Costeja González* [2014] EU:C:2014:317.

[78] See on the extraterritorial application of the GDPR: EDPB, Guidelines 3/2018 on the territorial scope of the GDPR (Art. 3), 2018; M Brkan, "Data Protection and Conflict-of-Laws: A Challenging Relationship" (2016) *European Data Protection Law Review* 324; DS Villa, "The Concept of Establishment and Data Protection Law" (2017) 4 *European Law Review* 491.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

Similarly, the NPDR applies in cases where the processing of electronic nonpersonal data in the EU is provided as a service to users residing or having an establishment in the EU, regardless of whether the service provider is established in the EU (Art. 2(1)(a) NPDR). However, since data portability under the NPDR is of a nonbinding nature, abiding is voluntary. As we suggest, a comprehensive data portability regime for personal and nonpersonal data would therefore be desirable at an international level.

3 Unilateral Power

The EU has been able to externalize its laws beyond its borders via the so-called unilateral power of the EU. While foreign firms are only bound by their national laws, they increasingly have been following EU data protection law.[79] This can, on the one hand, be explained by the advantages of international firms following a single rule, and preferring to harmonize their process and services for cost mitigation purposes.[80] In other words, it might be cheaper for a company to develop a single framework (that follows European data protection law), rather than two or more different ones (one following a stricter European regime, one a more lenient one). In the past, large technology companies like Facebook and Google have often made their data portability tools available to all their customers, independently of their location.[81] On the other hand, Apple took a staged approach and introduced its portability tool for users in Europe only in 2019 and made it available to US and Canadian users in 2020.[82] Apple, Facebook, Google, Microsoft, and Twitter are further contributing to the creation of an open-source framework connecting providers by translating provider specific APIs into common "data models" that can be transferred.[83]

4 International Soft Law

In the past, a number of guiding documents from the EU, such as the Article 29 Working Party Guidelines on the right to data portability in particular, already had a major impact on the interpretation of data portability concepts.[84] The guidelines, set by this former advisory board that has been replaced by the European Data Protection Board (EDPB) representing the data protection authorities of the EU

79 PM Schwartz, "Global Data Privacy: The EU Way" (2019) 94 *New York University Law Review* 778.
80 Ibid., further referring to the difficulties of firms to screen out EU customers.
81 "How Do I Download a Copy of Facebook?", https://perma.cc/7ULQ-AW3K; "Takeout", https://takeout.google.com/settings/takeout.
82 C Fisher, "Facebook Lets Users in the US and Canada Move Media to Google Photos" (Engadget, 30 April 2020), https://perma.cc/MRB5-7JKK.
83 "Data Transfer Project", https://perma.cc/PF9J-XL9L.
84 Guidelines on DP, note 47 above.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

member states, have been subject to extensive academic discussion and scrutiny by corporations.[85]

International soft law has long served as inspiration for national privacy codification, beginning with the OECD Privacy Guidelines of 1980, which were revised in 2013.[86] The Guidelines explicitly refer to personal data as an "increasingly valuable asset". Their aim has been to foster the free flow of information by preventing unjustified obstacles to economic development, namely by setting a minimum standard for national legal frameworks on privacy. The original 1980 OECD Privacy Guidelines were influential at first, encouraging the adoption of data protection laws in eight countries outside Europe (including Canada and Japan), but their impact diminished when the EU adopted its Data Protection Directive in 1995,[87] which went beyond the OECD Guidelines.[88] The OECD Guidelines also influenced the Asia-Pacific Economic Cooperation (APEC) Privacy Framework.[89]

As the OECD is reviewing its Guidelines, it could include a data portability norm in a future revision. However, as the OECD Guidelines only cover personal data, a right to data portability in the OECD Guidelines (alone) would not match its (optimal) scope.[90] Data portability should, in our view, rather be added to other international soft law instruments and encompass both personal and non-personal data.

5 International Hard Law

European portability concepts have been reflected in international treaties. This may be exemplified by the inclusion of a clause regarding the portability of telephone numbers in the Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP), a trade agreement between Australia, Brunei, Canada, Chile, Japan, Malaysia, Mexico, New Zealand, Peru, Singapore, and Vietnam.[91] As mentioned in subsection B, the right to telephone number portability in the former Art.

[85] See PN Yannella and O Kagan, "Analysis: Article 29 Working Party Guidelines on Automated Decision Making Under GDPR" (CyberAdviser, 16 January 2018), https://perma.cc/L34H-PNYQ. See for an overview on instruments of transnational exchange on data protection: F Zufall, Art 50 DSGVO Rn. 13, in Eßer/Kramer/von Lewinski (eds), *DSGVO/BDSG – Kommentar*, 7th ed. 2020.

[86] Recommendation of the Council concerning Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data – Annex [C(80)58/FINAL, as amended on 11 July 2013 by C (2013)79].

[87] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. OJ L 281, 23.11.1995, 31–50; G Greenleaf, "It's Nearly 2020, so What Fate Awaits the 1980 OECD Privacy Guidelines?" (2019) 159 *Privacy Laws & Business International Report* 18.

[88] Greenleaf, note 87 above.

[89] See "APEC Privacy Framework (2015)", https://perma.cc/Z6LT-57X5.

[90] The same applies for the APEC Privacy Framework of 2005.

[91] CPTPP, Article 13.5.4 ("Each Party shall ensure that suppliers of public telecommunications services in its territory provide number portability without impairment to quality and reliability, on a timely basis, and on reasonable and non-discriminatory terms and conditions").

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

30 Universal Services Directive[92] can be seen as a predecessor to data portability. Furthermore, the Eastern Caribbean Telecommunications Authority is planning to codify number portability in its new Electronic Communications Bill.[93]

Against this backdrop, the question arises whether international trade agreements should include data portability provisions going forward – either in competition chapters or in dedicated electronic commerce or digital trade chapters. Regulation on an international level, however, would require a supranational understanding of the modalities of data portability. Because data flows cross-countries, the need for a coherent regulation of portability is strong. Nonetheless, views on the modalities of a portability regime differ across states. The type of data it should cover, the concrete definition of "portability", the extent of interoperability required, the kinds of standardization of formats and interfaces, and whether retrieval "in a commonly used and machine-readable format" suffices are some of the many questions on which consensus should be reached. In this regard, the first experiences with the GDPR and NPDR will be crucial in determining the future of portability.

III CONCLUSION: TOWARD AN ALTERNATIVE CONCEPT OF DATA PORTABILITY

The EU regulations regarding data portability have an ambitious aim: to contribute to the creation of an effective data ecosystem characterized by the free flow of data. Both regimes, however, were designed to address very specific situations – the GDPR regime for users and their free-of-charge social media provider; the NPDR regime for business customers and their big data analytics providers. Both regimes find application beyond the use cases they were designed for. Instead of distinguishing between personal and non-personal data, a better regime for data portability could hinge on whether the value of generated data serves as compensation for the respective service providers' costs.

Ultimately, the distinction between personal and non-personal data can be challenged as inappropriate for data portability. Data portability is a concept that primarily serves the free flow of data rather than the protection of personal data. A classification distinguishing between raw and generated data has its advantages, particularly when it factors in the value of data. Competition law could rely more heavily on the value of data and its role in providing cost compensation, instead of using a terminology inherited from data protection law. Future data portability regimes may be better designed once they are removed from the realm of data protection. This assumes that the data subject is sufficiently protected by the remaining rights under the GDPR, especially via the right of access. Guaranteeing an effective right of access for raw and generated data is key.

[92] European Electronic Communications Code (Directive (EU) 2018/1972) OJ L 321, 17.12.2018, 36–214, Art. 106.

[93] "Electronic Communications Bill Revised 16 October 2019", https://perma.cc/AP3V-Z64E.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press

Consequently, we propose that the difference in choice of the regulatory regime for data portability should be made with a view to the value of data and depending on whether it provides compensation for cost-bearing. Raw data, being assigned to the customer or the data subject, would be portable, while generated data would require a more refined regime depending on whether it serves as a means of compensation.

https://doi.org/10.1017/9781108954006.012 Published online by Cambridge University Press