NBER WORKING PAPER SERIES

DATA, PRIVACY LAWS AND FIRM PRODUCTION: EVIDENCE FROM THE GDPR

Mert Demirer
Diego J. Jiménez Hernández
Dean Li
Sida Peng

Working Paper 32146
http://www.nber.org/papers/w32146

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2024, Revised December 2024

We thank Guy Aridor, James Brand, Alessandro Bonatti, Peter Cihon, Jean Pierre Dubé, Joe Doyle, Ben Edelman, Liran Einav, Sara Ellison, Maryam Farboodi, Chiara Farronato, Samuel Goldberg, Yizhou Jin, Garrett Johnson, Gaston Illanes, Markus Mobius, Christian Peukert, Devesh Raval, Dominik Rehse, Tobias Salz, Bryan Stuart, Taheya Tarannum, Joel Waldfogel, and Mike Whinston for helpful comments, and Abbie Natkin, Taegan Mullane, Doris Pan, Ryan Perry, and Bea Rivera for excellent research assistance. We are also grateful to Han Choi for copyediting assistance. We gratefully acknowledge the support of the National Institute on Aging, Grant Number T32-AG000186 (Li) and the National Science Foundation Graduate Research Fellowship under Grant No 214106 (Li). The views expressed herein are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Chicago, the Federal Reserve System, or the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w32146

Data, Privacy Laws and Firm Production: Evidence from the GDPR
Mert Demirer, Diego J. Jiménez Hernández, Dean Li, and Sida Peng
NBER Working Paper No. 32146
February 2024, Revised December 2024
JEL No. D22, L11, L51, L86

ABSTRACT

By regulating how firms collect and use data, privacy laws may alter firm demand for information technology inputs. We study how firms respond to privacy laws in the context of the EU's General Data Protection Regulation (GDPR) by using seven years of data from a global cloud-computing provider. Our difference-in-difference estimates indicate that, in response to the GDPR, EU firms decreased data storage by 26% and data processing by 15% relative to comparable US firms, becoming less "data-intensive." To estimate the costs of the GDPR for firms, we propose and estimate a production function where firms combine data and computation in firm production. We find that data and computation are strong complements and that firm responses are consistent with the GDPR representing a 20% increase in the cost of data. This increase translates into only a 0.1-0.5% rise in overall production costs because data plays a relatively small role in firm production compared to computation.

Mert Demirer
MIT Sloan School of Management
100 Main St
Cambridge, MA 02142
and NBER
mdemirer@mit.edu

Diego J. Jiménez Hernández
Federal Reserve Bank of Chicago
230 S LaSalle St
Chicago, IL 60604
diego.j.jimenez.h@gmail.com

Dean Li
Massachusetts Institute of Technology
deanli@mit.edu

Sida Peng
Microsoft
47 Brookline Street
Chestnut Hill, MA 02467
sidpeng@microsoft.com

A data appendix is available at http://www.nber.org/data-appendix/w32146

1 Introduction

In the information age, the production of goods and services increasingly relies on the processing of data (Agrawal et al., 2018; Goldfarb and Tucker, 2019). Since some of the most valuable data concerns personal information, its growing use has led to new policy attention and regulation. One of the most influential privacy laws is the European General Data Protection Regulation (GDPR), which was enacted in 2016 and affects more than 20 million firms across dozens of countries (GDPR.eu, 2019; Johnson, 2022). Many countries have since followed this example: as of early 2022, 157 countries had enacted legislation to secure data and privacy (Greenleaf, 2022).

While these privacy laws help harmonize and improve data collection practices, they can also be costly for firms (Peukert et al., 2022; Johnson et al., 2023; Aridor et al., 2023; Goldberg et al., 2023). For example, privacy regulations may generate a wedge between the marginal product of data and its (perceived) marginal cost, leading firms to substitute data with other inputs. Variations in these wedges across firms can result in misallocation of inputs in the economy (Hsieh and Klenow, 2009). Given the increasing role of data in firm production, understanding the cost of privacy regulations and how they affect firms' input decisions is of utmost importance.

However, large-scale empirical evidence of how privacy laws affect firm data decisions is scant. Studying this question is complicated for a number of reasons (Johnson, 2022). First, firms' data and computation usage are inherently difficult to observe, as standard firm datasets do not provide information on these measures. Second, there is no unified framework for analyzing the role of data in firm production (Veldkamp and Chung, 2023). Any such framework needs to be parsimonious while having enough flexibility to allow the effect of privacy laws to depend on the importance of data and computation for firms.

In this paper, we make progress on these fronts by studying how the GDPR affected firms' input choices by proposing a production framework with data and computation and using a dataset from a large global cloud-computing provider. The cloud is an ideal setting for this study because it enables us to observe firms' high-frequency data and computation usage across tens of thousands of firms over a seven-year period from 2015 to 2021. This data spans most major industries, from manufacturing to services, allowing us to analyze the effect of privacy regulations beyond the digital economy.

In our first set of analyses, we apply this data toward studying the direct effect of the GDPR on firm data and computation choices. We compare domestic firms in the European Union (EU) subject to the GDPR to similar, non-treated firms from the same industry in the US using a difference-in-differences approach. In our second set of analyses, we develop a production function model with data and computation. Using this model, we estimate how firms combine data and computation in production and quantify the wedges generated by the GDPR along with the corresponding increase in production costs.

We begin by summarizing the key features of the GDPR. The GDPR is a landmark privacy policy enacted in 2016 and implemented in 2018. Its regulations apply to all firms in the EU, as well as non-EU firms offering goods or services to "data subjects" within the EU. This law increased the cost of collecting and storing data for firms by requiring enhanced data protection, increasing penalties in case of data breaches, and giving consumers data-rights requests such as data correction and deletion. Survey evidence suggests that GDPR compliance is costly, ranging from \$1.7 million for small to medium-sized businesses to \$70 million for large ones (Accenture, 2018; Hughes and Saverice-Rohan, 2018).

Next, we discuss the specific context in which we observe firm data decisions: the cloud. Cloud computing is a widely adopted information technology (IT) that enables firms to store and process data remotely over the internet (Byrne et al., 2018; Brand et al., 2024). Using data from our cloud computing provider, we observe firm-level monthly usage of "storage"—the amount of data stored in gigabytes—and "compute"—the number of core-hours of computation. We also observe other information, such as prices and the locations of the data centers where firms do computation and store their data. We match our cloud usage data to other data sources that provide information on firm characteristics.

Our first set of results comes from an event study design comparing data and computation use among comparable firms in the EU and the US after the GDPR. We find that EU firms stored on average 26% less data than US firms two years after the GDPR. The direction of this relative decline in data is perhaps unsurprising, given that the GDPR primarily regulates data usage, but the magnitude is noteworthy. We also find that EU firms decreased their computation relative to US firms by 15%—implying that firms became less data-intensive after the GDPR.[1] Furthermore, our heterogeneity analysis suggests that these patterns are present across all industries we study (software, services, and manufacturing). Finally, we look at how these effects vary with a measure of regulatory stringency across EU countries created by Johnson (2022), as enforcement of the GDPR is delegated to individual countries. Although the differences are not statistically significant at the 5% level, our estimates suggest a larger decline in both data and computation in countries with higher regulatory stringency.

While our event study findings provide direct evidence of the GDPR's impact on firms' data and computation inputs, they offer a limited understanding of the associated

[1] It is ex-ante unclear how the GDPR would affect computation; this effect theoretically depends on the substitutability between data and computation (Acemoglu, 2002).

economic costs of the regulation. Since data are inputs in firm production, recovering the regulatory wedges from firms' input choices and, ultimately, the effect of regulation on firms' overall production costs depends on how firms use data in production.

Motivated by this, we propose a production function model where firms aggregate data and computation through a constant elasticity of substitution (CES) function. This aggregation function, which we call "information production," includes two key parameters: (i) *the firm-specific compute-augmenting productivity*, which determines relative factor intensity of computation and data (Doraszelski and Jaumandreu, 2018; Raval, 2019; Demirer, 2022) and (ii) *the elasticity of substitution between computation and data*, which governs how firms adjust these inputs in response to changes in factor prices (Hicks, 1932). Our model accommodates many of the uses of data proposed in the literature (Jones and Tonetti, 2020; Farboodi and Veldkamp, 2022) and emphasizes the role of computation in firm production.

Our production function model provides an input demand function that links firms' cost-minimizing data and computation choices to input prices and model parameters. Using a shift-share design to instrument input prices, we estimate this input demand function for each industry to recover the parameters of the production function. We find that data and computation are strong complements in production, with elasticity of substitution ranging from 0.44 (services) to 0.34 (manufacturing). This complementarity suggests that firms cannot easily substitute toward computation when faced with increased data costs. To our knowledge, this is the first estimate of a production function with data inputs, which contributes to our understanding of production functions in modern firms.

To recover the distortion generated by the GDPR, we model it as a wedge between the variable cost of storing data in the cloud and the total variable cost that includes GDPR compliance costs. This wedge arises from various sources of regulatory costs, including penalties in case of breaches, higher data security requirements, and the need for detailed data records. We estimate firm-specific wedges by attributing them to the changes in post-GDPR input choices unexplained by changes in input prices or changes in the elasticity of substitution.

Our estimates suggest that the GDPR increased the variable cost of data inputs by 20% for firms on average. Firms in data-intensive industries faced higher costs, with the largest effect observed in the software sector (24%), followed by manufacturing (18%) and services (18%). What determines the increase in costs? To provide suggestive evidence, we analyze the relationship between firm-specific wedges and two firm characteristics: (i) firm size, measured by the number of employees, and (ii) compute-augmenting productivity, estimated from the production function. We find that larger and more compute-intensive firms experienced smaller cost increases from the GDPR.

In the last part of the paper, we use our production function estimates to quantify how the 20% GDPR-induced increase in the cost of data translates into firms' total variable production costs. Our analysis proceeds in two steps. First, we analyze its impact on the cost of aggregating data and computation in information production and then examine how changes in information costs affect total production costs.

We find that although the average firm-level wedge is quite large (20%), the resulting increases in the variable cost of producing information are quite low (3.7%), primarily because data's cost share in information production is considerably smaller than that of computation (19% vs. 81%). In other words, although strong complementarity limits firms' ability to substitute data for computation when data becomes more costly, the expenditure share of the data is small to begin with, limiting the GDPR's impact on the cost of information.

Next, to estimate the effect of the GDPR on the total production cost, we perform a simple back-of-the-envelope calculation, assuming a CES technology in IT and non-IT inputs (e.g., capital, labor). We calibrate this model using estimates from Lashkari et al. (2024) and other data sources. We find that the GDPR increases variable production costs by 0.47% for software firms, with smaller effects in less data-intensive industries. When aggregated across all EU firms in the industries we analyze, this corresponds to an annual increase in production costs of approximately €16 billion.

We conduct additional analyses to show that our results are robust to many concerns. First, we show that our results are similar when we exclude multi-cloud firms, suggesting that results are not driven by EU firms substituting toward other cloud providers. Second, we find similar results when estimating our empirical strategy using only start-ups, which tend to use cloud computing as their only IT—suggesting that substitution to on-premises IT (hybrid cloud) is not a large concern. Third, we show that our results are not driven by differential trends in cloud prices in the EU and the US. Finally, we estimate our specification while excluding firms using web services, showing that the results do not only come from websites, which experienced cookie consent changes under the GDPR.

Nevertheless, we acknowledge some relevant limitations of our study. Unlike many previous GDPR studies, our paper is based on a large sample of firms. While this allows us to draw more generalizable conclusions about firms' data uses, the trade-off is that we observe less detailed information than an in-depth single-firm study. For example, although we observe detailed measures of the quantity of information stored in our data, we cannot be as precise about the role of data for the firm as more focused studies can be. Finally, we highlight that our results focus on the costs imposed on firms and do not speak to the consumer benefits of privacy (Arrieta-Ibarra et al., 2018), where further evidence is needed to understand the benefits of these laws.[2]

Contribution to the Literature The first body of literature we contribute to is the research on the impact of the GDPR on firms (Johnson, 2022). This literature finds that the GDPR decreased the investment in technology ventures, encouraged app exit, and discouraged app development (Kircher and Foerderer, 2020; Jia et al., 2021; Janßen et al., 2021). Several papers document adverse impacts on digital tracking and advertising: the GDPR decreased the usage of tracking technology tools (Lefrere et al., 2022; Aridor et al., 2023; Lukic et al., 2023), decreased page views and e-commerce revenue (Goldberg et al., 2023), decreased the number of website visits (Schmitt et al., 2022), increased market concentration in the advertising sector (Peukert et al., 2022; Johnson et al., 2023) and increased search frictions (Zhao et al., 2021). On the benefits side, Aridor et al. (2023) find an increase in the average value of data for advertising, while Godinho de Matos and Adjerid (2022) document improvements in targeting effectiveness due to the GDPR. Although most evidence suggests that the GDPR has impacted data-driven economic activity, Zhuo et al. (2021) find a null short-term effect on the formation and termination of internet infrastructures between GDPR and non-GDPR countries.[3]

While our paper builds on an identification strategy similar to some of these GDPR papers, it differs in two aspects. First, because of the unique feature of our data, we go beyond digital outcomes to analyze firms' data and computation decisions, margins directly targeted by the regulation. By studying these outcomes, we also complement the literature that focuses on accounting and aggregate measures of firm performance, such as profit and sales (Koski and Valmari, 2020; Frey and Presidente, 2024). Second, we take a production function approach. This approach allows us to structurally estimate the role of data and computation in production and to calculate the cost of the GDPR for firms.

Second, our study relates to the production function literature by estimating a production function with data inputs (Olley and Pakes, 1996; Ackerberg et al., 2015). A recent theoretical literature has proposed different ways of how firms use data, with Jones and Tonetti (2020) modeling data as a non-rival input generated as a byproduct of production and Farboodi and Veldkamp (2022) modeling data as a productivity-enhancing input through better prediction. On the empirical side, some studies have included IT in firm production by using various IT expenditures, such as software and hardware, as inputs (Brynjolfsson and Hitt, 2003; Lashkari et al., 2024). We contribute to this literature by estimating a micro-level production function that incorporates physical measures of two fundamental modern IT inputs: data and computation.

[2] Quantifying the privacy benefits is known to be difficult (Acquisti et al., 2016; Lin and Strulov-Shlain, 2023).
[3] A recent literature has studied the California Consumer Privacy Act (Canayaz et al., 2022; Doerr et al., 2023).

Third, our paper is related to the misallocation literature, which studies inefficiencies in factor allocations resulting from various frictions (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). We employ a similar empirical strategy by modeling distortions as wedges between the marginal revenue product of an input and its price. Although this literature often abstracted from the origins of frictions, some recent papers have focused on their sources, such as labor market institutions (Bertrand et al., 2021), market power (Asker et al., 2019; Peters, 2020) and monopsony power (Berger et al., 2022). We contribute to this literature by studying the input distortions introduced by a landmark global regulation.

Our paper also contributes to the economics of privacy literature (Goldfarb and Tucker, 2011, 2012; Acquisti et al., 2016; Athey et al., 2017; Choi et al., 2019; Montes et al., 2019; Ichihashi, 2020; Loertscher and Marx, 2020; Chen et al., 2021; Krähmer and Strausz, 2023) by evaluating the effects of the largest privacy regulation on firms.

2 Institutional Setting

This section first discusses the relevant details of the GDPR. We then describe cloud computing technology, the setting for our primary data source in this paper.

2.1 The European General Data Protection Regulation

There is perhaps no policy more important in the modern privacy landscape than the GDPR. As Johnson (2022) notes, "In many ways, the GDPR set the privacy regulation agenda globally." As such, understanding the consequences of the GDPR is vital not only because of its impacts on firms but also because of its crucial role in shaping privacy laws. In this section, we describe the key features of this policy and its implications for firms.

The GDPR is a set of rules that govern the collection, use, and storage of personal data belonging to individuals within the EU. It was enacted in April 2016 and came into force in May 2018. By consolidating and enhancing existing privacy provisions, the GDPR introduced a harmonized approach to privacy regulations across the EU.[4] We provide a detailed description of the changes required for firms after the GDPR in Appendix B.1 and summarize its most important characteristics below.

The GDPR applies whenever the firm that controls the data ("data controller") is established in the EU or whenever the individuals ("data subjects") whose data is collected are located in the EU, regardless of their citizenship or residence (Article 3). It broadly defines personal data as any information relating to an identified or identifiable natural

[4]Unlike the GDPR, which is directly binding across the EU, the preceding Directive 95/46/EC had to be incorporated into each member state's national laws, leading to variation in its implementation across states.

person (Article 4). This includes information such as name, address, email address, and internet protocol (IP) address. It applies to *all* personnel data both in the client and employee context, making even business-to-business firms subject to compliance.

Two aspects of the GDPR are particularly important for our paper. First, the GDPR takes a data protection approach rather than a consumer protection approach as in the US (Boyne, 2018; Jones and Kaminski, 2020). A data protection approach imposes a set of costly responsibilities on firms to protect data, in addition to a substantive system of individual rights. Second, the GDPR takes a risk-based approach to data protection without clarity on the specific measures firms must take, making implementation firm-dependent (Hustinx, 2013; Gellert, 2018). For example, Article 25 (Data Protection by Design and by Default) uses phrases such as "taking into account the state of the art, the cost of implementation [...] as well as the risks" and requires that controllers "implement appropriate technical and organizational measures [...] in an effective manner." This risk-based approach makes regulatory costs heterogeneous across firms.

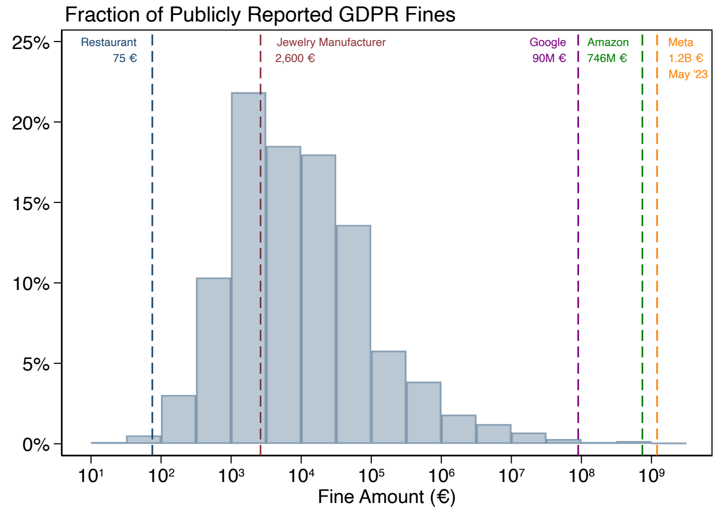
From the firm perspective, the GDPR mainly increased the cost of collecting and storing data by imposing costly responsibilities on firms. These include designating a data protection officer (Article 37), preparing data protection impact assessments (Article 35), implementing appropriate technical and organizational measures for data security (Article 32), keeping a record of processing activities (Article 30), providing timely notifications in case of data breaches (Article 33), fulfilling consumers' requests for data transfer, erasure, or rectification (Article 14-21), and paying penalties in case of data breaches (Article 83).[5]

The cost of complying with the GDPR can vary depending on the size and complexity of an organization. There are no official statistics, but most survey evidence suggests that complying with the GDPR is costly. The estimates range from an average of \$3 million (Hughes and Saverice-Rohan, 2018) and \$5.5 million (Ponemon Institute, 2017) to \$13.2 million (Ponemon Institute, 2019) depending on the composition of surveyed firms. The survey evidence indicates that a large percentage of the costs (between one-fifth and one-half) are labor costs, followed by technology, outside consulting, and internal training (Ponemon Institute, 2019; Hughes and Saverice-Rohan, 2019).

The changes mandated by the GDPR entail both fixed and variable costs. For example, the cost of having a data protection officer may not scale with data size, so it could be considered a fixed cost. On the other hand, the costs of handling customers' access requests, the liability in case of a data breach, and keeping data secure would increase

[5]Firms also must have a legal basis for processing personal data. Contrary to popular belief, consent is not the only appropriate legal basis—contractual necessity, legal obligation, vital interests, public task, and legitimate business interest may also serve as a basis for processing data (Article 6).

Figure 1: Distribution of Publicly Reported GDPR Fines



Notes: The figure presents the distribution of 1,730 publicly available GDPR fines, noting that not all GDPR fines are made public. Appendix B.3 describes the data collection process. Fines are presented in undeflated nominal terms (€), and five examples from the data have been highlighted.

with data and firm size. As such, it may be more sensible to interpret these kinds of costs as marginal costs. We provide a detailed classification of GDPR costs into these fixed and variable cost categories and present corresponding survey evidence in Appendix B.2.

In addition to these direct costs, firms may also incur indirect costs such as cybersecurity insurance or penalties if they are found to be non-compliant.[6] Non-compliant firms may face fines of up to 4% of an organization's annual *global* revenue or €20 million (whichever is greater). We scraped publicly available GDPR fine data from a database maintained by CMS, an international law firm.[7] In Figure 1, we provide the size distribution of these GDPR fines.[8] We note two key features of these fines. First, enforcement is not limited to large violations: 25% of the fines have been under €2,000 levied on small businesses. Second, the GDPR applies to a much broader set of businesses and industries than just software and technology firms. Figure 1 highlights some of these cases, which include fines on restaurants and manufacturers.

2.2 Our Setting: Cloud Technology

Cloud computing provides scalable IT resources on demand over the internet. According to the National Institute of Standards and Technology (Mell et al., 2011), cloud computing

[6] There are likely additional costs beyond the direct financial costs of compliance, including opportunity costs of diverting existing employees towards GDPR compliance and disruption caused by operational changes.

[7] See `https://www.enforcementtracker.com`. Appendix B.3 describes this dataset.

[8] The total cumulative fines imposed in this dataset have amounted to over €3 billion, with over 1,700 being fined. This figure is likely to be an underestimate because not all GDPR fines are made publicly available.

is defined as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released." Cloud computing has experienced rapid growth since its introduction, with nearly 80% of firms using at least one cloud function as of 2018 (Zolas et al., 2021).[9]

We focus on the two primary cloud services: storage and computation. Storage services allow users to store data in a data center. Computation services allow users to run applications and perform computations in a virtual machine (VM). Firms may use storage and computing services in multiple parts of their production, such as powering digital services, optimizing logistics, supporting product development, and handling administrative tasks like human resources and accounting. Firms may also use storage without computing services, such as a newspaper hosting website photographs in the cloud and providing them directly without computing. However, it is rare to observe firms using computation without storage, although non-data simulations might serve as an exception.[10]

From the researchers' point of view, the cloud's existence and ubiquity provide important advantages over traditional IT. Because cloud computing is typically provided by large third-party firms, it is possible to aggregate data from tens of thousands of firms. Moreover, cloud providers keep detailed records of their users' activity for billing purposes, allowing usage to be tracked consistently over time.

Despite these advantages, there are limitations to using data from cloud computing. First, many firms use a mix of cloud computing and on-premises IT, especially during the transition to the cloud. In such cases, we can only observe firm data in the cloud and not from their on-site hardware, which may bias our results if the GDPR changes the composition of cloud and on-site data. Second, firms frequently use cloud services from multiple providers, known as multi-cloud (Accenture, 2022). For these firms, a decline in cloud usage from one provider could come from substitution to another provider. We take these concerns seriously and provide several robustness checks in our empirical strategy.

3 Data and Summary Statistics

This section describes the main datasets used in the paper and presents basic summary statistics. We leave the exact data construction details to Appendix C.

[9] See Jin and McElheran (2017); Jin (2022); DeStefano et al. (2023) for recent studies on firm's cloud adoption and the effects of cloud technology on firms.

[10] See several case studies of how firms in different industries use cloud computing at AWS Case Studies, Azure Customer Stories, and Google Cloud Customers. All web links in the paper were accessed on Nov 26, 2024.

3.1 Cloud Computing Data (2015-2021)

We obtain information through one of the largest cloud technology providers. Using this data, we observe monthly-level storage and computation usage information for the universe of their customers between 2015 and 2021. For each service, we observe the number of units purchased, the location of the data center, the date, and the price paid. Therefore, we have both the physical unit of usage and expenditures.

We measure storage in gigabytes and computing in core-hours (number of cores × number of hours). Core-hours are a commonly used metric to quantify computational work in cloud computing.[11] We use this data to construct monthly usage at the firm-location (data center) level for storage and computation from July 2015 to December 2021. As a result, we can observe data stored in the US and EU separately by the same firm. Through this data, we also observe SIC industry codes, headquarters location, and whether a firm is a start-up or not.[12] Additional details on this data are provided in Appendix C.1.

One limitation of our dataset is that it does not allow us to see which specific data firms are collecting nor the exact ways in which they use the data. This limits our ability to speak to some important questions about how firms specifically use data.

3.2 Cloud Computing Usage from Other Providers (2016-2021)

To address the concern of observing data from a single provider, we use an establishment-level IT data panel produced by a market research company called Aberdeen (previously known as "Harte Hanks"). Aberdeen compiles data on cloud technology adoption (including provider) using web crawling, surveys, and publicly available sources. This dataset covers around 1.9 million companies worldwide between 2016 and 2021 at the yearly level. Previous versions of this data have been widely used by researchers to construct measures of IT usage.[13] We use this data to identify single cloud firms and examine differential changes in market shares of cloud providers in the EU and US around the GDPR.

3.3 Other Datasets: Firm Characteristics

Aberdeen also provides information on other firm characteristics, such as employment and revenue from Duns & Bradstreet. We match our cloud computing data to Aberdeen

[11] To illustrate the concept, consider the example of a software engineer in a startup who runs a VM with eight cores for five hours. In this case, the usage is recorded as 40 units of compute.

[12] The "start-up" classification is defined internally by the cloud technology provider.

[13] See e.g., Bloom et al. (2012). Note that Aberdeen's data has undergone changes in recent years, relying more on web scraping and extrapolation than on surveys. We conduct cross-checks with our internal data to assess the quality of Aberdeen's accuracy for cloud adoption. See Appendix C.3 for more details.

Table 1: GDPR Applicability Matrix by Location from Peukert et al. (2022)

		Firm Location	
		EU	US
Location of Consumer / Employee Data Used	EU	**Case 1** GDPR applies *Art. 3(1) GDPR*	**Case 3** GDPR applies *Art. 3(2) GDPR*
	US	**Case 2** GDPR applies *Art. 3(1) GDPR*	**Case 4** GDPR does not apply –

Notes: Table is taken from Table 1 of Peukert et al. (2022). The matrix shows whether the GDPR is applicable to firms located within and outside the EU.

firms using a matching procedure described in Appendix C.3 based on name, location, domain, and other information. We match close to 60% of our cloud firms to the Aberdeen dataset. We further augment our data by merging it with employment data from the European Orbis database from Bureau van Dijk through name and domain matching. With this procedure, we link cross-sectional employment data to approximately 80% of the European firms. We use the employment information in 2018 to define firm size.

3.4 Sample Construction and Summary Statistics

We begin by presenting a framework that will allow us to classify firms by their exposure to the GDPR. Following Section 2, Table 1 presents information on whether the GDPR applies to firms depending on the location of the firm and data subjects (using the language from Peukert et al., 2022). Now, while we cannot directly observe the location of each firm's employees and consumers, we use the fact that we can observe firm server locations to approximate the locations of their consumers and employees. We view this as a reasonable approximation because firms tend to choose data centers close to them to reduce latency (Greenstein and Fang, 2020). We argue that firms based solely in one geographic region are unlikely to use servers across the Atlantic unless they have consumers or employees located in the other location.[14]

By combining information on the locations of firm server choices before the GDPR with the locations of firm headquarters, we attempt to categorize firms into the four cases described in Table 1. We consider a firm multi-national (Cases 2 and 3) if they use data centers both in Europe and in the US. We consider a firm to be a domestic EU or US firm

[14]One piece of evidence that supports server location choice being predictive of firm location is that when we construct EU vs US firms classifications using only server locations, the assigned regions coincide with the headquarter locations in our data for 98% of the firms.

Table 2: Summary Statistics

Industry	Number of Firms	Share Compute (%)	Share Storage (%)	Mean Storage	Mean Compute	Mean Data Intensity	Share EU (%)
Services	15,886	36.3	31.9	844	628	1.84	40.9
Software	9,480	17.6	20.8	690	670	1.69	59.8
Manufacturing	3,095	10.5	11.6	1,293	986	1.81	54.4
Retail Trade	2,152	5.2	5.4	1,101	917	2.02	46.9
Finance & Insurance	2,057	11.4	10.8	1,652	1,571	1.89	44.9
Wholesale Trade	1,945	3.7	4.5	925	885	2.10	52.3
Other	2,689	15.3	15.0	1,714	1,616	2.23	46.1
All	37,304	100.0	100.0	1,000	803	1.86	48.1

Notes: Table presents summary statistics from our matched sample of firms. A description of the sample's construction can be found in Section 3.1, and a more detailed description of the sample construction can be found in Appendix C. Industries are defined as the ten divisions classified by SIC codes, with the exception of software firms, which are carved out of the services division and represent SIC codes 7370 - 7377. For confidentiality purposes, mean storage and compute have both been normalized such that mean storage is denoted by 1,000 units. We calculate mean data intensity at the firm level while restricting to firms that use both storage and computing services.

(Cases 1 and 4) if they use data centers only in Europe or in the US.[15] As we explain later in the paper, our empirical strategy focuses on comparing domestic EU and US firms, and therefore, these domestic firms constitute our main sample throughout the paper.[16]

As we discuss in Appendix C.2, we restrict our attention to firms that continuously used our cloud provider's services for the full year beginning two years prior to the introduction of the GDPR. This sample accounts for 90% of storage and computation. We use this sample restriction to focus our analysis on relatively stable users of cloud computing. Our sample is, therefore, comprised of firms that are both responsible for the vast majority of storage and computation in the pre-GDPR period and that have been continuously attached to our cloud computing provider.

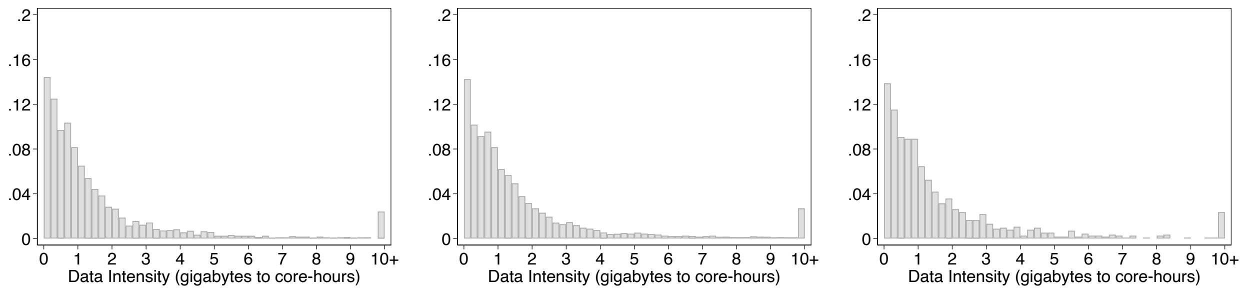
Table 2 presents summary statistics for our baseline sample of nearly forty thousand firms. We categorize each firm's industry by using the firm's SIC code, and we intentionally split software firms from other firms in the services division due to their large share in our sample.[17] Therefore, throughout the paper, we use "services" to describe firms in the service industry excluding software firms, and "software" to describe firms in the software industry. The majority of firms belong to the services (43%) and software (25%) industries, but firms from manufacturing and various other industries are also represented in our

[15] We also include UK firms in our EU sample. The UK was part of the EU when the GDPR came into effect on May 25, 2018. After the UK's withdrawal from the EU, the GDPR was incorporated into UK law as the UK GDPR, which largely mirrors the provisions of the GDPR, with some minor changes.

[16] While multinational firms are important, their exposure and responses to the GDPR are more complex than those of domestic firms, which requires us to focus on domestic firms.

[17] We define software firms as those with SIC codes between 7370 and 7377.

Figure 2: Histogram of Data Intensity by Industry



(a) Software Firms **(b)** Services Firms **(c)** Manufacturing Firms

Notes: Figure presents a histogram of data intensity at the firm level, defined as the ratio of data stored to computation (the ratio of gigabytes to core hours) for each industry. Industries are defined through SIC codes (with the exception of software firms, which are carved out of the services division). We limit the sample to firms that have ever used both storage and computation ($N = 11,858$).

sample. As reported in Columns 5-6, while there is variation in usage across industries—likely driven in part by the difference in the average size of firms using cloud computing—we observe significant storage and computation in all industries. We also note some slight variation in the share of firms in the US versus the EU by industry in Column 8, although each region always accounts for at least 40% of firms in each industry.

Lastly, Column 7 of Table 2 presents the mean data intensity for each industry, which is defined as the ratio of storage to computation. We find that the average data intensity ranges from 1.69 to 2.23. However, these averages mask significant within-industry heterogeneity, as shown in Figure 2, which plots the distribution of data intensity for the three largest industries in our sample. The large firm-level variation in data intensity suggests that the roles of data and computation likely vary across firms.[18] This result is consistent with the large evidence of within-industry heterogeneity in other firm outcomes, such as productivity (Syverson, 2011), labor shares (Kehrig and Vincent, 2021), and markups (Autor et al., 2020; De Loecker et al., 2020). As we will see in Section 5, taking into account this heterogeneity will be important when developing a production function framework with data and computation.

4 Event Study Evidence

In this section, we apply an event study design to study the effect of the GDPR on firms' data and computation decisions. We begin by defining our empirical strategy and providing intuition for our identifying assumptions. Next, we present our baseline estimates and

[18] This result remains even if we focus on more narrowly defined 4-digit SIC industries.

discuss the robustness of our strategy across various alternative samples and specifications.

4.1 Empirical Strategy

Our empirical strategy aims to identify the causal effect of the GDPR on firms' computation and data choices. In order to identify a relevant treatment and control group for our strategy, we turn to our classifications of firm locations from Section 3. Following Table 1, we define "Case 1" as our treatment group and "Case 4" as our control group.

Notably, these two definitions exclude multi-national firms (i.e., those with branches and/or consumers across countries). We choose to do so for two reasons. First, we may think of multi-national firms as being partially treated: only some of their data may be subject to the GDPR. Thus, we might want to separate the estimation of the treatment effects of these firms from the fully treated firms (Case 1). Second, multi-national firms may systematically differ from the control firms that we define (Case 4). They may respond to the GDPR along different margins than our control group, choosing to shift data, computation, and even business operations into or out of the EU—responses that are outside the scope of this paper.

We focus on three outcomes: data, computation, and "data intensity" (the ratio of data to computation). Our empirical specification uses a difference-in-differences design and estimates the following regression:

$$\log(Y_{it}) = \sum_{q \neq -1} \beta_q \cdot \mathbb{1}_{\{\mathrm{EU}_i\}} + \alpha_i + \tau_{kqs} + \varepsilon_{it}, \tag{1}$$

where Y_{it} is the outcome of interest for firm i, in month t. We use q to denote quarter, k to denote industry, and s to pre-GDPR cloud usage decile. In this specification, α_i is a firm-level fixed effect while τ_{kqs} are industry-by-quarter-by-size-decile fixed effects which allow for time trends to differ flexibly in each quarter for an industry-size decile combination.[19] We define eleven industries using the ten mutually exclusive and exhaustive divisions defined by one-digit SIC codes and carving out software from services.

We estimate this specification for the sample period from July 2015 to March 2020.[20] The coefficients of interest, β_q, represent the difference in outcomes relative to the quarter before the GDPR came into force. The identifying assumption of our empirical strategy is a

[19]We measure cloud usage deciles for storage and computation outcomes by using a firm's computation or storage, respectively, as measured one year before the GDPR. For data intensity, we use terciles of firm storage interacted with terciles of firm compute to increase power.

[20]Even though we have data for a few more quarters, we end the sample in March 2020 to rule out the effects of the COVID-19 pandemic. This sample restriction also limits the potential effects of another privacy law, the California Consumer Privacy Act, which came into effect in January 2020.

conditional parallel trends assumption. We take advantage of our large sample and allow time trends in our outcomes to vary flexibly by industry and initial cloud usage levels in our baseline specification, with 110 distinct bins for each quarter (11 defined industries × ten pre-GDPR cloud usage deciles).

To discuss the short- and long-run estimates of the effect of the GDPR, we also present results in a table format using an alternative regression specification given by:

$$\log(Y_{it}) = \delta_1 \cdot \mathbb{1}_{\{\mathrm{EU}_i\}} \cdot \mathbb{1}_{\{t \in \mathrm{Jun/18\text{-}May/19}\}} + \delta_2 \cdot \mathbb{1}_{\{\mathrm{EU}_i\}} \cdot \mathbb{1}_{\{t \in \mathrm{Jun/19\text{-}May/20}\}} + \alpha_i + \tau_{kqs} + \varepsilon_{it}, \quad (2)$$

where the notation of α_i and τ_{kqs} is the same as in Equation (1). Our estimates are relative to the excluded group, which is the pre-GDPR period. Thus, the short-run coefficient (δ_1) and long-run coefficient (δ_2) estimate the average difference in Y_{it} between treated and untreated firms in the first and second year after the GDPR came into force.

4.2 Results

Our main event study results are shown in Figure 3, which plots the estimated coefficients β_q from Equation (1) for three outcomes. We discuss each of these outcomes separately and present the corresponding short- and long-run estimates from Equation (2) in Table 3.

Results on Data Panel (a) of Figure 3 shows the results for data storage. First, we find no evidence of significant differential pre-GDPR trends in the US and EU, as all pre-GDPR coefficients are close to zero. We also find limited evidence for anticipation effects, which is consistent with the survey evidence that only 10% of firms expected to be compliant with the GDPR before May 2018 (Ponemon Institute, 2018). After the implementation of the GDPR, however, firms in the EU, relative to US firms, started to decrease their relative amount of data stored gradually, with cumulative effects growing steadily over the two years after the GDPR. The fact that the decrease is gradual rather than sudden may be due to the fact that it took time for firms to implement necessary changes, as noted by Aridor et al. (2023) in the case of a large website.

The decline in data is perhaps not surprising, as the GDPR increased the cost of storing data. What is perhaps more surprising, however, is the magnitude of the effect. Table 3 shows that the short-run effect is around a 13% decrease in data while the long-run effect doubles to around 26%.[21]

[21] Importantly, firms are not necessarily deleting data, as our identification strategy relies on comparing EU and US firms. Data storage for EU and US firms could be increasing but at different rates.

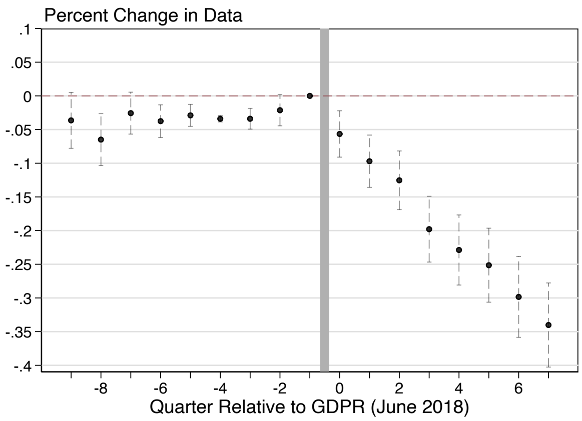
Table 3: Short- and Long-Run Effects of the GDPR
(Data, Computation, and Data Intensity)

	(1)	(2)	(3)	(4)
	Panel A. Dependent variable: Log of Data			
Short-Run Effect	-0.129	-0.132	-0.125	-0.134
	(0.018)	(0.017)	(0.017)	(0.017)
Long-Run Effect	-0.257	-0.260	-0.228	-0.242
	(0.024)	(0.024)	(0.024)	(0.024)
Observations	1,143,149	1,143,149	1,143,149	1,143,149
US Firms	16,409	16,409	16,409	16,409
EU Firms	16,281	16,281	16,281	16,281
	Panel B. Dependent variable: Log of Computation			
Short-Run Effect	-0.078	-0.082	-0.132	-0.148
	(0.016)	(0.016)	(0.016)	(0.016)
Long-Run Effect	-0.154	-0.164	-0.224	-0.256
	(0.024)	(0.024)	(0.024)	(0.024)
Observations	672,942	672,942	672,942	672,942
US Firms	10,294	10,294	10,294	10,294
EU Firms	8,927	8,927	8,927	8,927
	Panel C. Dependent variable: Log of Data Intensity			
Short-Run Effect	-0.072	-0.071	-0.025	-0.021
	(0.020)	(0.020)	(0.020)	(0.019)
Long-Run Effect	-0.131	-0.126	-0.049	-0.035
	(0.029)	(0.029)	(0.029)	(0.029)
Observations	418,803	418,803	418,803	418,803
US Firms	5,487	5,487	5,487	5,487
EU Firms	5,872	5,872	5,872	5,872
Time Trends Vary By:	Industry × Pre-GDPR Size Deciles	Pre-GDPR Size Deciles	Industry	-

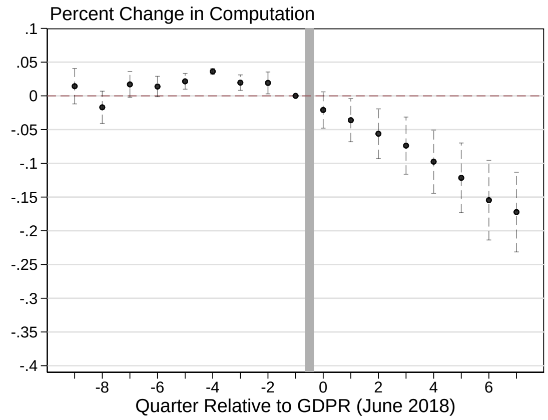
Notes: Table presents estimates of Equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the effect of the GDPR in the first and second year after the GDPR came into force. Column (1) presents our baseline specification, where we allow for time trends to vary flexibly across industry and pre-industry size decile interactions. Column (2) restricts these time trends so that they only vary by pre-GDPR size decile, while Column (3) only allows for variation at the industry level. Column (4) shows estimates when we include no time-trend interactions. Industries are defined as the ten divisions classified by SIC codes. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we define "size decile" as the interaction between storage and computation terciles when measured in the period. Standard errors are clustered at the firm level.

Figure 3: Event Study Estimates of the Effects of the GDPR on Cloud Inputs

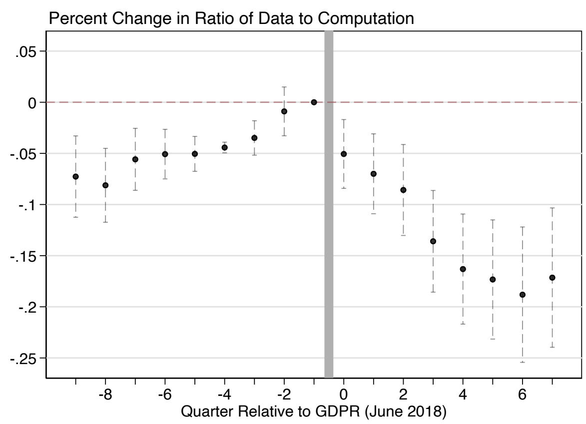
(a) Effect on Data



(b) Effect on Computation



(c) Effect on Data Intensity



Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. Dotted bars represent the 95% confidence intervals, and standard errors are clustered at the firm level. Sample sizes are presented in Table 3.

Results on Computation Turning towards computation, we first note that there is no clear theoretical prediction for how the GDPR should affect firms' computation decisions. The GDPR's primary goal is to protect personal data, with limited direct implications for computation. Therefore, the effect of the GDPR on computation likely depends on the elasticity of substitution between computation and data and the intensity of these inputs in the production function. If data and computation are substitutes, firms can respond to increases in data costs by substituting away from data toward computation. On the other hand, if data and computation are complements, then an increase in data cost would lead to a decrease in computation. Thus, the direction and magnitude of firm computation responses is ultimately an empirical question.

Panel (b) of Figure 3 shows that EU firms gradually decreased their computation relative to US firms after the introduction of the GDPR. The effect on computation is smaller than what we observe for data, with only a 15% decline two years after the GDPR. Similar to the results on data, we find no evidence of significant differential pre-GDPR trends.

The results on computation are also important because they indicate that firms do not simply eliminate (or stop accumulating) unused data. One potential explanation for our data results is that before the GDPR, firms stored data that they never utilized and subsequently deleted it to comply with the GDPR. Our findings suggest that this hypothesis is unlikely because of the substantial reduction in computation, which we conjecture would not have happened if data that was not being used was simply eliminated.[22]

Results on Data Intensity Comparisons of the magnitudes between our data storage and computation results suggest that firms became less data-intensive after the GDPR. However, in order to account for potential compositional effects, we investigate the effects of the GDPR on data intensity by using the natural logarithm of the ratio of storage to computation as an outcome. We estimate our specification on firms that used both types of inputs for the full year beginning exactly two years before the GDPR came into force.

Panel (c) of Figure 3 shows that firm data intensity decreased immediately after the GDPR. Panel (c) of Table 3 estimates a decrease of around 7% in the short run and 13% in the long run. The fact that firms in the EU become less data-intensive post-GDPR (relative to comparable US firms) suggests that storage and computation are likely complements in production, which we revisit using a production framework in Section 5.[23]

Robustness of Results There are several potential threats to our identification strategy. In Appendix D, we go through these threats and provide evidence indicating they do not drive our results. We summarize the main exercises below, and we leave the additional exercises (e.g., alternative sample definitions and empirical specifications) in Appendix D.

The most salient identification threat is that we observe only one, albeit large, cloud provider. What we observe as declines in cloud usage could simply be firms substituting usage towards other providers ("multi-cloud") or to their on-premises IT services ("hybrid cloud"). For multi-cloud, we show that our results are similar when we restrict our sample to firms that only use our cloud provider according to Aberdeen data (Table OA-2 and Figure OA-8). For hybrid cloud, we first show that our empirical exercise yields similar

[22] This hypothesis also appears unlikely because cloud computing incurs a marginal cost for storing data, even if it remains unused. Additionally, in Section 5, we find that firms are responsive to changes in cloud prices.

[23] Table 3 also shows the robustness of including flexible time trends by industry and size-decile fixed effects. We observe that excluding the pre-GDPR size fixed effects results in similar storage estimates, slightly higher (in absolute value) computation estimates, and lower data intensity estimates. These differences likely reflect compositional variations in treatment responses by firm size between EU and US firms.

results for the start-up firms in our sample, which are less likely to use on-premises IT (Table OA-4 and Figure OA-10).[24] Second, we find no evidence of differential trends in interest in hybrid cloud usage—as proxied for by Google Trends—across the EU and the US.[25] Third, we show that EU firms are less likely to leave the cloud relative to the US firms after the GDPR (Figure OA-13). Therefore, it is unlikely that the declines we observe are simply driven by substitution to on-premises IT.

Another natural explanation for our results is the possibility of differential price trends in the EU and the US. If cloud prices increase in the EU relative to the US post-GDPR (perhaps to cover GDPR compliance costs, for example), we could see a decline in data and computation even without the GDPR having any additional effects on firms. To check this hypothesis, we use the paid prices for cloud storage as a dependent variable and find no differential price changes between the EU and the US (Figure OA-12).

We also consider whether our results are particularly being driven by websites' cookie consent notices and the clauses governing the collection and storage of data from websites. We might expect firms with active website use—which we proxy for through the usage of cloud-based web services—to be more affected by the policy than those without. Table OA-5 shows larger effects among firms that used web services in storage and computation. However, we find that the data and computation adjustments of web users and non-web users are proportional and that their reductions in data intensity are similar.

4.3 Heterogeneity

By Industry The relationship between data and computation may vary by industry, depending on how each industry incorporates data inputs into its production processes. For this reason, we investigate whether the effects of the GDPR on data and computation vary across four mutually exclusive and exhaustive industry groups: software, services, manufacturing, and all other industries. Table 4 shows our estimates of the short- and long-run effects of the GDPR when we estimate Equation (2) across different industry groups.[26] One striking result is the breadth of our results: we find declines in data, computation, and data intensity across all industry groups. This suggests that the direct impact of the GDPR extends beyond the subset of previously studied industries or mechanisms—e.g.,

[24] See Jin and McElheran (2017) and Ewens et al. (2018) for research supporting this assumption.

[25] Figure OA-11(a) shows no differential time trends in hybrid cloud-related searches between the US, the UK, or Germany, which is suggestive that differential uptake of hybrid cloud services in the EU is unlikely to explain our results. Furthermore, hybrid cloud remains an order of magnitude less popular as a search term than cloud computing (Figure OA-11(b)). See Appendix D.1 for more information.

[26] We show the quarterly dynamics in Figures OA-1 and OA-2, and the (lack of) pretrends at the industry level.

Table 4: Short- and Long-Run Effects of the GDPR
(Heterogeneous Effects by Industry)

	Baseline	Software	Services	Manufacturing	Other Industries
	(1)	(2)	(3)	(4)	(5)
Panel A. Dependent variable: Log of Data					
Short-Run Effect	-0.129	-0.113	-0.080	-0.259	-0.190
	(0.018)	(0.035)	(0.026)	(0.063)	(0.037)
Long-Run Effect	-0.257	-0.253	-0.180	-0.404	-0.354
	(0.024)	(0.048)	(0.036)	(0.086)	(0.051)
Observations	1,143,149	291,781	486,457	94,612	270,299
US Firms	16,409	3,196	8,141	1,141	3,931
EU Firms	16,281	5,150	5,912	1,508	3,711
Panel B. Dependent variable: Log of Computation					
Short-Run Effect	-0.078	-0.078	-0.048	-0.171	-0.077
	(0.016)	(0.032)	(0.024)	(0.051)	(0.033)
Long-Run Effect	-0.154	-0.150	-0.100	-0.322	-0.163
	(0.024)	(0.050)	(0.037)	(0.073)	(0.049)
Observations	672,942	165,752	270,846	65,532	170,812
US Firms	10,294	2,050	4,623	900	2,721
EU Firms	8,927	2,747	3,204	914	2,062
Panel C. Dependent variable: Log of Data Intensity					
Short-Run Effect	-0.072	-0.084	-0.084	-0.078	-0.043
	(0.020)	(0.042)	(0.031)	(0.066)	(0.039)
Long-Run Effect	-0.131	-0.196	-0.161	-0.043	-0.069
	(0.029)	(0.064)	(0.045)	(0.097)	(0.055)
Observations	418,804	103,606	168,020	41,449	105,729
US Firms	5,487	1,054	2,473	496	1,464
EU Firms	5,872	1,755	2,123	610	1,384

Notes: Table presents estimates of Equation (2) of δ_1 and δ_2, re-estimated across for various industry divisions. For comparison, Column (1) presents our baseline estimates across all industry divisions. Column (2) restricts our sample to software firms, which are defined through SIC codes 7370 - 7377. Column (3) restricts the sample to services firms, Column (4) restricts the sample to firms in the manufacturing division, and Column (5) presents estimates on the remaining firms in the sample (non-software, non-services, and non-manufacturing industry divisions). Standard errors are clustered at the firm level.

websites or venture capital investments—to affect firms across all industries.[27]

Furthermore, we find substantial heterogeneity between industries in the magnitudes of the effects. Panel A shows that the most significant decreases in data in response to the GDPR come from manufacturing firms (40% in the long run), followed by software firms (25%), and services firms (18%). Similarly, Panel B shows that for computation, the fall is largest in magnitude for manufacturing (32% in the long run), followed by software (15%) and services (10%).

While it may seem surprising that IT-intensive industries like software and services firms have more muted responses to the GDPR than manufacturers, this may reflect several factors. First, manufacturers are still subject to the GDPR if they sell directly to customers, employ workers in the EU, or work with EU suppliers or trading partners. Second, manufacturers might be able to substitute compute and data with other inputs more easily in response to the GDPR because data and computation are less essential parts of their production than software and services firms. Alternatively, they might be less sophisticated in terms of their existing data infrastructure and comply with the GDPR by simply reducing data usage.[28]

Finally, Panel C of Table 4 shows results for data intensity. We find that data intensity decreases in all industries, although the standard errors are large for some of these estimates. Our point estimates suggest that long-run data intensity decreases the most in the industries that experienced the smallest declines in storage and computation.

By Regulatory Stringency Although the GDPR harmonized data protection regulations across the EU, enforcement was delegated to each country's data protection authority. Thus, enforcement policies can vary across countries due to differences in resources available to data protection authorities and their approaches to data protection. (Johnson, 2022). Because of these differences, we might expect firms in countries with more lenient regulators to respond less to the GDPR. To test this hypothesis, we use a measure of perceived regulatory strictness created by Johnson et al. (2023) using data from European Commission (2008) that varies at the country level. This measure calculates a z-score for each country based on firms' stated perceptions of their country's relative data protection regulatory strictness. We then classify each firm as above or below the normalized median strictness in the survey according to the strictness of their country's regulator.

We modify Equation (2) by introducing two additional coefficients to account for poten-

[27]Some papers in the literature find a decrease in venture capital investment in the EU after the GDPR (Jia et al., 2021; Janßen et al., 2021), but our results extend to both large firms and industries, such as manufacturing, that are less like to receive venture capital investment.

[28]For some commercial products offered to manufacturers for GDPR compliance, see GrowthDot and Ground-Labs. For an overview of how GDPR applies to manufacturers, see Data Protection Laws for Manufacturers.

Table 5: Short- and Long-Run Effects of the GDPR by Regulatory Strictness

	Data (1)	Computation (2)	Data Intensity (3)
Short-Run Effect	-0.028	-0.061	-0.042
	(0.044)	(0.032)	(0.042)
Long-Run Effect	-0.040	-0.047	-0.015
	(0.055)	(0.049)	(0.059)
Observations	1,143,149	672,942	418,803
EU Firms	16,281	8,927	5,872

Notes: Table presents estimates of Equation (2) with an additional term to measure the effect of above-median GDPR strictness. The short-run term captures the triple interaction of the short-run post-GDPR coefficient, the EU categorical variable, and a categorical variable indicating firms in the above-median regulatory stringency countries. The long-run term repeats the same procedure but uses the long-run post-GDPR period instead. Regulatory strictness is measured according to Johnson et al. (2023) using data from European Commission (2008). For data intensity, we define "size decile" as the interaction between data and computation terciles when measured in the period. Standard errors are clustered at the firm level.

tial heterogeneity by regulatory stringency. Specifically, we interact a categorical variable indicating above-median stringency with the EU categorical variable to measure the short- and long-run differences in Y_{it} for EU firms across different levels of regulatory stringency.

Table 5 summarizes these results. The interaction coefficients—although many are not statistically significant—are all negative, suggesting that firms in above-median regulatory strictness countries face larger declines in data, computation, and data intensity. In the short run, data decreases by 2.8 pp. more in above-median strictness countries than in below-median ones, while computation declines by 6.1 pp. more. In the long run, data and computation go down by 4 pp. and 4.7 pp. more in above-median strictness countries, respectively. Similarly, data intensity shows larger declines for firms in above-median strictness countries. Overall, these findings suggest a non-negligible role for regulatory stringency beyond the simple presence of privacy regulation itself.

4.4 Discussion

Our findings suggest that EU firms responded to the GDPR by storing less data, performing less computation, and becoming less data-intensive compared to US firms. These results provide direct and large-scale evidence that firms comply with the GDPR by adjusting their data inputs. Moreover, the results are not driven by a single industry or websites affected by cookie consent policies, indicating the far-reaching implications of the GDPR.

Although these findings provide insight into how privacy laws affect firm behavior, they do not offer a comprehensive understanding of the economic costs imposed on firms. Such an analysis requires understanding the role of data in firm production and considering firms' adjustment margins in response to privacy regulations. For this reason, we take a more structural approach in the next section.

5 A Model of Production with Data

This section introduces a production function framework with data and computation and estimates its parameters. We use our framework to study both how firms use data and computation in production and how privacy regulations might affect these decisions. We model the GDPR as a wedge between the cost of storing data and the total (perceived) cost of data that includes regulatory costs. We focus on estimating the size of this wedge, corresponding increases in production costs, and their implications for firms.

5.1 Production Function with Data

Firm i in month t produce output Y_{it} by combining compute (C_{it}), data (D_{it}) and other inputs (X_{it}):

$$Y_{it} = F\big(X_{it}, I_{it}(C_{it}, D_{it}), \omega_{it}\big),$$

where the function $I_{it}(\cdot)$ aggregates compute and data inputs to be used in firm production and ω_{it} is firm productivity. It is natural to model the contribution of data and computation to firm production in this way, as these inputs are inherently interdependent: firms use computation to process data, and the processed data is then combined with other inputs. We assume a CES functional form for the aggregation of data and computation:

$$I_{it}(C_{it}, D_{it}) = \big(\omega^c_{it}(C_{it})^\rho + \alpha D^\rho_{it}\big)^{1/\rho}, \tag{3}$$

where ω^c_{it} is compute-augmenting productivity, α denotes data intensity, and $\sigma = 1/(1-\rho)$ is the elasticity of substitution between data and compute.[29] While CES imposes parametric assumptions, it offers flexibility in the elasticity of substitution, the key parameter that governs how firms re-optimize their inputs in response to the GDPR. We refer to the intermediate input $I_{it} = I_{it}(C_{it}, D_{it})$ as "information" throughout our analysis.[30]

[29] As we will show later, α can be normalized without loss of generality because the ratio of ω^c_{it} to α serves as a sufficient statistic that determines the relative intensity of compute and data in production. We retain α in the notation to highlight the role of data intensity in the derivations we will present later.

[30] Information I_{it} lacks a natural unit in the production function shown in Equation (3). This is because any monotone transformation $h(\cdot)$ of the information production function can be offset by applying $h^{-1}(I)$ inside the F function. However, as we show in Appendix E.4, this scale invariance does not affect our identification strategy of wedges and associated production cost increases as we focus on changes in firms' costs due to the GDPR instead of its levels. This robustness to monotone transformations also accommodates information

Our empirical analysis will primarily use the CES model of data and compute aggregation in Equation (3) rather than the full production function. This choice is motivated by the lack of a standardized framework for modeling data in firm production. For example, data could increase overall firm productivity (Jones and Tonetti, 2020), serve as an input in production (Bessen et al., 2022), increase labor productivity (Agrawal et al., 2019), and increase revenue by better customer targeting (Eeckhout and Veldkamp, 2022).[31] While this limits potential counterfactual analyses we could perform, we consider it a reasonable trade-off given our study's large-scale coverage across firms and industries.

Our production function model in Equation (3) includes a firm-specific compute-augmenting productivity term, ω_{it}^c, to capture heterogeneity in production technology across firms. This heterogeneity can arise from two main sources. First, firms may differ in their inherent production technologies regarding how much data they need, making production more data-intensive for some firms than others. Second, even with similar underlying technologies, firms may achieve different levels of compute productivity through differences in resources they have, including technical infrastructure and human capital (e.g., advanced software tools and skilled engineers). The large firm-level variation in data intensity that we documented in Figure 2 underscores the importance of accounting for these technological differences.

Our approach relies on estimating input demand functions derived from the CES form under the assumption that firms choose inputs to minimize static production cost of I_{it}, taking ω_{it}^c as given. In particular, we assume that C_{it} and D_{it} are variable inputs that firms optimize every period. We view this assumption as reasonable for cloud computing given that firms can easily adjust their usage of storage and computation. We also assume that firms are price-takers in the input markets for C_{it} and D_{it}, which is again a reasonable assumption for cloud computing because firms typically pay a linear price by the hour.[32] Overall, this static cost minimization assumption enables us to bypass the dynamics of firms' decision-making, which would necessitate additional assumptions.

We use p_{it}^c and p_{it}^d to denote the input prices for compute and data, which may vary across firms, as we explain later. Based on the cost minimization assumption, we derive the following first-order condition (FOC) for firms' ratio of compute and data choices from

production to be non-constant returns to scale through the transformation $h = I^{\theta}$. However, we note that our empirical strategy does not identify the returns to scale parameter.

[31] More formally, our setting covers: (i) $Y = f(X)\omega(I)$ (productivity increasing), (ii) $Y = f(X, I, \omega)$ (input in production), (iii) $Y = f(X, \omega^L(I) \cdot L, \omega)$ (labor-augmenting), and (iv) $R = p(I)f(X, \omega)$ (price discrimination). In these examples, Y and R are output and revenue; ω^L is labor-augmenting productivity, and p is the output price. In each specification, information affects a different part of the production function.

[32] An exception is very large firms, which can negotiate their prices bilaterally. Since we focus on domestic firms, this exception likely affects a very small fraction of our sample. See footnote 34 for more information.

the CES production function:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma \log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma \log(\omega^c_{it}), \tag{4}$$

where $\gamma = -\sigma \log(\alpha)$. We provide the complete derivations in Appendix E.1 and also show in Appendix E.2 that we obtain the same FOC if we were to include labor (e.g., software engineers) in the information production function. We note that the level of ω^c_{it} is not separately identified in this equation from α, so we normalize it to 1 in the estimation.

According to this FOC, the relationship between input ratios and input prices is governed by the elasticity of substitution. A key feature of this equation is that the elasticity of substitution and compute-augmenting productivity can be estimated from firms' input demand functions alone, without requiring data on other inputs or outputs. This property of the CES functional form has been commonly used in the literature for estimating the elasticity of substitution (Doraszelski and Jaumandreu, 2018; Raval, 2019; Demirer, 2022).

Although our framework extends the production function literature by incorporating computation and data, it has certain limitations. While we account for potential variations in data quality across firms through ω^c_{it}, we assume that data is homogeneous within each firm. This limitation becomes particularly relevant if, for instance, firms have data types with varying levels of quality, and the GDPR impacts the composition of data. Relaxing this assumption requires incorporating different data types into the production function, which we do not observe. It is worth noting, however, that the assumption of homogeneous inputs within a firm is a common practice in the production function literature.

5.2 The GDPR as a Cost Shock to Data

We model the GDPR as a cost shock to data input—as we have extensively argued data is the main focus of GDPR regulations. While some aspects of the GDPR do pertain to computation (e.g., Article 30, records of processing activities), the effects of the regulation on data are significantly larger, and computation is less salient to regulators than data.

As mentioned before in Section 2 and in Appendix B.2, the GDPR increased the fixed and variable costs of using data. For example, customer data-rights requests under the GDPR may impose variable costs on firms that increase with data amount. Similarly, the probability of a data breach and penalties in case of non-compliance likely increase with the amount of data firms have.[33] By contrast, fixed costs are one-time expenses that do not vary with data amount—e.g., hiring data protection officers and developing a

[33]This observation aligns with the fact that larger firms tend to receive more substantial fines in our fine data.

data protection management system. Since fixed costs do not affect input demand in the intensive margin, we focus on modeling the variable cost.

We make the following assumptions about data costs before and after the GDPR:

$$\textbf{Pre-GDPR:} \quad \tilde{p}^d_{it} = p^d_{it}, \qquad\qquad \textbf{Post-GDPR:} \quad \tilde{p}^d_{it} = (1+\lambda_i)p^d_{it}.$$

Here, p^d_{it} represents the variable cost of data without the GDPR (i.e., the cost of storing data paid to the cloud provider), and $\tilde{p}^d_{it}$ is the cost after accounting for the regulatory costs introduced by the GDPR. Therefore, λ_i denotes the wedge between the actual cost of data and the total variable cost that includes complying with the GDPR. We follow the literature and model λ_i as a multiplicative wedge (e.g., Chari et al., 2007; Hsieh and Klenow, 2009). This wedge is firm-specific because compliance costs are likely to be heterogeneous across firms, depending on their size and the types of data they collect. Alternatively, we can also interpret λ_i as each firm's perceived cost of the GDPR, as they may hold different beliefs about enforcement or have varying levels of risk aversion that affect the expected cost of liability in the event of a data breach (Ganglmair et al., 2024).

Although we have modeled the GDPR as affecting the variable cost of data, we show in Section E.3 that our estimation procedure is consistent with several other interpretations of the GDPR. First, we show that if there are other unobserved variable costs to data that generate wedges before the GDPR, our estimate captures the additional wedges driven by the GDPR. Next, we consider a model with data-augmenting productivity where the GDPR generates a negative shock to this productivity by reducing the effectiveness of data. We demonstrate that our estimation procedure approximately recovers the size of this negative productivity shock in such a model. Finally, if the GDPR generates wedges in compute in addition to data, our wedge estimate will reflect the ratio of data-to-compute wedges, making our estimate of costs conservative.

5.3 Identification of Parameters

Our end goal is to estimate the production function parameters and the firm-level wedges introduced by the GDPR. To illustrate the potential identification problems when estimating these objects, consider the FOC in Equation (4) after the GDPR for EU firms:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma \log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma \log(1+\lambda_i) + \sigma \log(\omega^c_{it}). \tag{5}$$

This FOC reveals a fundamental identification challenge: the GDPR wedge, λ_i, cannot be separately identified from a level shift in ω^c_{it} post-GDPR. Intuitively, firms may decrease their compute-to-data ratio either because their compute-augmenting productivity has increased or because the GDPR has imposed additional data costs. Without additional assumptions, we cannot distinguish these two cases using our data. To this end, we impose the assumption that compute-augmenting productivity can be decomposed as follows:

$$\log(\omega_{it}^c) = \log(\omega_i^c) + \log(\phi_t^c) + \log(\eta_{it}). \tag{6}$$

Here ω_{it}^c is decomposed into a firm-specific component (ω_i^c), an industry-specific time trend (ϕ_t^c), and a mean-zero (in logs) idiosyncratic component (η_{it}). This decomposition suggests that we need to control for ω_i^c and ϕ_t^c to identify the wedges generated by the GDPR.

Our identification strategy therefore involves two steps. In the first step, we recover ω_i^c and ϕ_t^c using data from EU firms in the pre-GDPR period and data from US firms, respectively. In particular, we assume that firm-specific compute technology does not change after the GDPR and that each EU industry follows the same compute-technology trend as the same industry in the US. With these assumptions, we can control for firm-specific compute-augmenting technology in the second step and estimate the GDPR wedge as a percentage of the observed data storage cost. We explain each of these steps below and provide more detail in Appendix F.4.

5.3.1 First Step: Identification of Compute Productivity and Elasticity of Substitution

To estimate the elasticity of substitution and compute-augmenting productivity, we use pre-GDPR data and estimate the following equation:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma_1 + \sigma_1 \log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \sigma_1 \log(\omega_i^c) + \sigma_1 \log(\phi_t^c) + \sigma_1 \log(\eta_{it}), \tag{7}$$

where σ_1 is the pre-GDPR elasticity of substitution. Two important considerations arise when estimating this equation. First, the estimation requires variation in the data-to-compute price ratio across firms over time. Second, these prices might be correlated with unobservable productivity shocks (η_{it}). To address this endogeneity, it is important to understand the factors generating price variation in cloud computing.

Cloud computing prices typically vary depending on the region where the data center is located. These posted prices can be considered orthogonal to the firm-level idiosyncratic compute-augmenting productivity shocks (η_{it}) because it is unlikely that any single firm is large enough to affect them conditional on ϕ_t^c. In addition, cost improvements and increased competition were the main drivers of price changes in the last decade (Byrne et al., 2018). However, the prices that firms pay may differ from these posted prices for two reasons. First, firms may have differential preferences over data center locations based on distance. Second, firms may receive a percentage discount from the listed price based on long-term commitments.[34] These two sources of price variation can create endogeneity because, for example, firms with a high compute-augmenting productivity shock may switch between data centers based on price differences, or they may receive long-term commitment discounts. We address these potential sources of endogeneity by developing a shift-share design (Bartik, 1991; Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022).

We first introduce the broad intuition behind our instrument. Our approach aims to address endogeneity in prices by leveraging two features of our data. First, because we observe both list prices and paid prices, we can use changes in list prices to instrument for the changes in paid prices. These changes, however, are still predictive of the prices that firms face because discounts are applied to list prices. Second, we construct a measure of exposure to specific data centers for each firm and period. We use historical exposure shares rather than contemporary ones because previous data center choices are sunk. However, previous data center choices remain predictive of current data centers firms use due to switching costs, as transferring data between locations is time-consuming and costly, especially for large datasets. As a result, firms' data center location choices are highly persistent over time.

More formally, the shift-share design combines list prices with variations in firms' pre-existing data center location choices. We construct instruments z_{it}^{d} and z_{it}^{c} for the data storage and computation prices each firm i faces at time t. The exposure shares in a given period are calculated as the ratio of firm i's usage in a specific data center to its total usage across all data centers. These exposures yield the following equation for the instrument:

$$z_{it}^{\{c,d\}} = \sum_{l \in \mathscr{L}} s_{il(t-12)}^{\{c,d\}} p_{lt}^{\{c,d\}} \tag{8}$$

where $s_{il(t-12)}^{\{c,d\}}$ denotes firm i's usage share for data center location l as measured 12 months

[34] Cloud providers offer discounts if firms commit to using cloud resources over a specific period of time (typically one or three years). These discounts are called "reserved instance" or "committed use" discounts, depending on the provider. These discounts are applied to the list prices and are the same across customers except for very large customers, who might individually negotiate prices. A survey of 750 companies conducted in 2023 finds that only one-third of them use these discounts (Flexera, 2023), which is likely lower during our sample period and among domestic firms. Moreover, firms that receive long-term commitment discounts can resell or refund their commitments for a fee in most major cloud providers (AWS Reserved Instance Marketplace). Therefore, we believe that linear prices are a good approximation for firms' monthly decisions of storage and computation. For examples of these pricing policies, see AWS Reserved Instance Market and Azure Reserved VM Instances. We provide more information about cloud computing pricing in Appendix F.1.

before t, $p_{lt}^{\{c,d\}}$ is the price index for each service in location l at time t, and $\mathscr{L}$ denotes the set of data center locations.[35] Our exposure shares are lagged by 12 months because contemporaneous shares are susceptible to reverse causality. While shift-share instruments can be driven by assumptions about either the exogeneity of "shares" or the independence and exogeneity of "shocks" (Borusyak et al., 2022), the identifying assumption underlying our exposure shares is most similar to the "shares" assumption discussed in Goldsmith-Pinkham et al. (2020). In particular, the exclusion restriction behind our shift-share design is that contemporary shocks to the compute-augmenting productivity of each firm are exogenous to the changes in the ratio of list prices of cloud computing in the firm's historical data center choices, controlling for industry-specific trends.[36]

We use z_{it}^c/z_{it}^d as an instrument for price ratio p_{it}^d/p_{it}^c and estimate Equation (7) for three EU industries (software, services, and manufacturing) separately using pre-GDPR data, as the pre-GDPR data does not include a regulatory wedge. This allows us to estimate compute-augmenting productivity (ω_i^c) and elasticity of substitution parameters before the GDPR. We also estimate Equation (7) for US industries over the entire sample period, as US firms do not experience any regulatory distortion. This allows us to recover the industry-specific compute-augmenting productivity trends, ϕ_t^c, for US industries.

5.3.2 Second Step: Identification of the Cost of the GDPR

In the second step, we use the EU post-GDPR data to estimate the wedges generated by the GDPR and the EU post-GDPR elasticity of substitution between compute and data. Incorporating this into the firm's input demand function, we obtain the following equation:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma_2 + \sigma_2 \log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \sigma_2 \log(1+\lambda_i) + \sigma_2 \log(\omega_i^c) + \sigma_2 \log(\phi_t^c) + \sigma_2 \log(\eta_{it}), \quad (9)$$

where σ_2 is the post-GDPR elasticity of substitution. Here, unlike in the pre-GDPR period, GDPR wedge, λ_i affects the compute-to-data ratio: a higher λ_i leads firms to substitute away from data toward compute. To use this equation for identifying λ_i, we make the following assumptions:

Assumption 1. *Firms' compute-augmenting productivity (ω_i^c) remains the same after the GDPR.*

[35] We provide more detail on our price index construction in Appendix F.2.

[36] One example of a potential threat to identification would be if η_{it} are correlated over time after accounting for aggregate industry time trends, and this caused firms to select data centers with specific trends in the ratio of prices. However, given that our model is estimated with the ratio of prices rather than direct price levels and considering that forecasting data center-specific trends in these price ratios is difficult, we view our identification assumption as reasonable for the setting. We provide further details for the instrumental variable construction in Appendix F.3.

We note that this assumption still allows for industry-specific trends in compute due to $\log(\phi_t^c)$ in Equation (9). The assumption also does not restrict firms' ability to respond to the GDPR by changing their compute-to-data ratio. Rather, it implies that the firm-specific component of the underlying information production technology remains the same.

At this point, it is worth comparing our approach to the approaches taken in the literature that estimate distortionary wedges. The large literature on misallocation identifies distortions as the difference between the marginal product of an input and its price (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). This literature assumes that production technology is the same across firms up to Hicks-neutral productivity because otherwise, the firm-specific wedges cannot be distinguished from arbitrary firm-level heterogeneity in production technology. We face the same identification problem but take a different approach. Instead of assuming homogeneous production technology, we allow for heterogeneity through compute-augmenting productivity but assume that this heterogeneity is time-invariant within a window of a few years. We note that both approaches have strengths and weaknesses, but we believe that in our context, it is essential to allow for heterogeneous production technology.

We also differ from the misallocation literature by analyzing input demand functions for two variable inputs—one distorted and one undistorted—instead of estimating a full production function. In our approach, we can net out the sources of distortions common to both inputs, such as market power, and recover the distortion specific to the data input. This identification strategy is similar to the approach used in the literature to identify input market power from the two variable inputs (Morlacco, 2020; Kirov and Traina, 2023).

Assumption 2. *EU and US industries follow the same time trends in aggregate compute-augmenting productivity (ϕ_t^c) post-GDPR.*

This is the second assumption necessary for identifying the GDPR wedges by controlling for industry-level changes in compute-augmenting productivity. Otherwise, any level shift in the compute-to-data ratio of EU firms post-GDPR may be attributed to arbitrary changes in aggregate compute-augmenting productivity in the EU. Therefore, we use the estimated post-GDPR industry trend from the US firms to control for industry trends in the EU. The parallel trends we find within industries before the GDPR in our reduced-form results support this assumption (Figures OA-1, OA-2 and OA-3).

With these two assumptions, we can estimate the following equation:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma_2 + \sigma_2\Big(\log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \log(\hat{\phi}_t^c)\Big) + \sigma_2\Big(\log(1+\lambda_i) + \log(\hat{\omega}_i^c)\Big) + \log(\eta_{it}), \tag{10}$$

Table 6: Elasticity of Substitution Results by Industry

Industry	Software		Services		Manufacturing	
	OLS	IV	OLS	IV	OLS	IV
Elasticity of Substitution (σ_1)	0.45	0.41	0.45	0.44	0.38	0.34
	(0.02)	(0.03)	(0.02)	(0.04)	(0.04)	(0.05)
First-Stage (Instrument)	-	0.15	-	0.16	-	0.18
	-	(0.01)	-	(0.01)	-	(0.01)
Firm FE	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
F-Stat	-	5,637	-	5,147	-	1,949
Observations	130,560	130,560	106,594	106,594	44,708	44,708

Notes: Table presents our estimation results of the elasticity of substitution between data and compute across industries. Estimates are presented for pre-GDPR elasticities for EU firms (σ_1). Standard errors are calculated using 100 bootstrap repetitions at the firm level.

where $\hat{\omega}_i^c$ denotes estimates of compute-augmenting productivity using pre-GDPR data and $\hat{\phi}_t^c$ denotes the estimates of compute-augmenting productivity trend of the US firms. By estimating this equation using EU firms' post-GDPR data, we can identify our main object of interest (λ_i) along with the post-GDPR elasticity of substitution.[37] Our specification, therefore, allows for changes in the elasticity of substitution post-GDPR. To account for the uncertainty in the two-step estimation procedure, we calculate standard errors via a bootstrap procedure that treats firms as independent observations and resamples firms with replacement within industries over 100 repetitions. We provide the details of the estimation procedure in Appendix F.

6 Production Function Estimation Results

This section provides results on the elasticity of substitution between data and compute, the wedges introduced by the GDPR, and how these wedges affect firms' production costs.

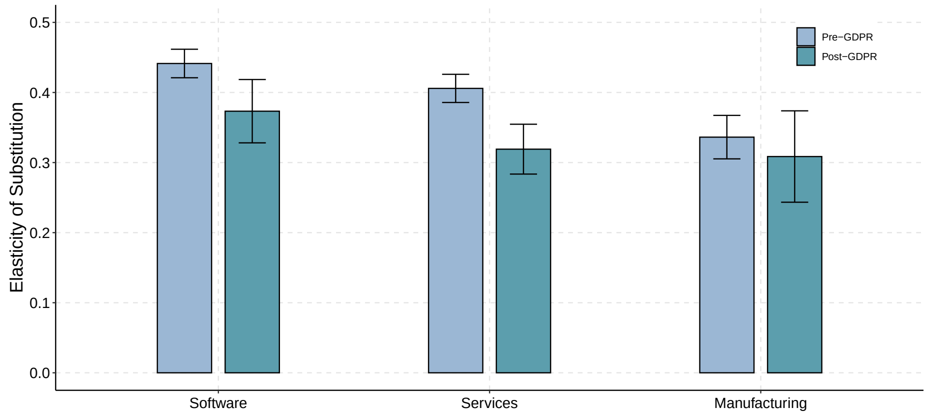
6.1 The Elasticity of Substitution Between Data and Computation

We begin by presenting estimates for the elasticity of substitution using pre-GDPR data. Table 6 presents these elasticities for three industries separately—software, services, and manufacturing—using both OLS and IV estimates.[38] We also present the first-stage es-

[37] Appendix F.5 provides useful intuition behind the identification of λ_i. Roughly speaking, the estimated wedges capture the variation in data intensity (the ratio between inputs) among comparable EU and US firms that is not explained by changes in prices, changes (over time or across regions) in the elasticity of substitution, or differences in compute-augmenting productivity.

[38] We exclude "other industries" analyzed in the event study from the production function analysis because we do not want to impose a single production function for different industries.

Figure 4: Elasticity of Substitution Between Data and Compute for EU Firms



Notes: Figure presents our estimation results of the elasticity of substitution between data and compute (σ) across industries, and we present separate estimates for the pre- and post-GDPR (σ_1 and σ_2, respectively). Solid lines denote the 95% confidence intervals, and standard errors are calculated using 100 bootstrap repetitions at the firm level.

timates for each industry and their associated F-statistics. The first-stage coefficients are positive, indicating a positive relationship between our shift-share instruments and the contemporaneous prices faced by firms. Our results also indicate F-statistics in the thousands, suggesting that our instruments strongly correlate with the endogenous variables.

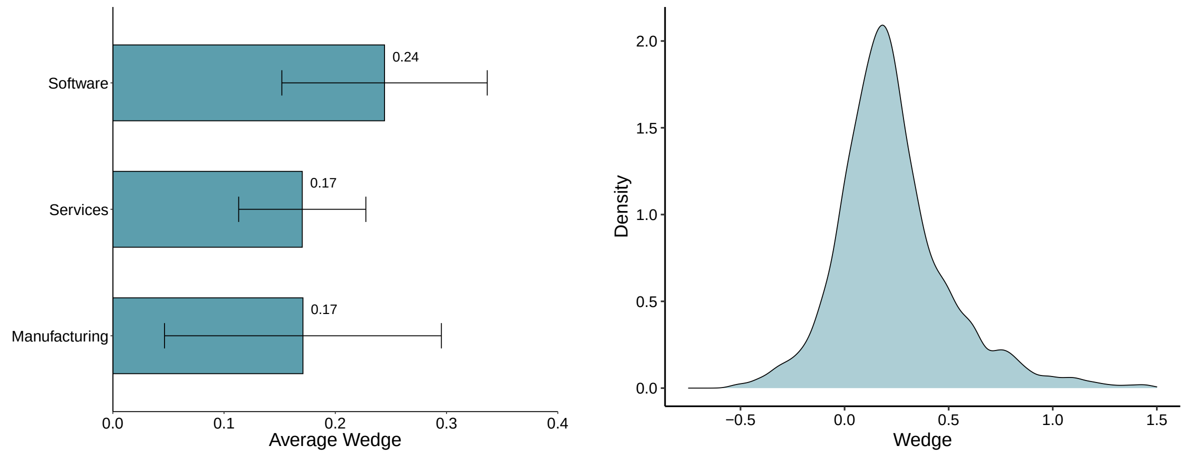
Our estimates suggest that data and compute are strong complements in all industries, with the estimated elasticities ranging from 0.34 to 0.44. The larger magnitudes in the software industry suggest that software firms can more easily substitute between data and compute. Furthermore, our IV estimates are smaller than the OLS ones. This bias is consistent with our intuition that firms with higher compute-augmenting productivity may be more likely to search for lower relative computation prices.

We also assess whether the GDPR led to any change in production technology in Figure 4, which separately reports the elasticity of substitution estimates before and after the GDPR for EU firms. We find a slight decline in the elasticity of substitution in all industries, suggesting that even though the GDPR affected firms' production technology, its impact is limited.[39]

Finally, we compare our estimated elasticity of substitution between data and compute to the existing elasticity of substitution estimates of other inputs to understand how the IT inputs differ from traditional inputs. While the estimates vary, they range from 0.3 to 0.7 for capital and labor (Caballero et al., 1995; Chirinko, 2008; Raval, 2019) and from 1.5 to 3 for labor and intermediate inputs such as materials (Chan, 2023; Peter and Ruane, 2023).

[39]In Figure OA-4, we repeat this exercise for US firms for comparison. We find comparable elasticities of substitution for firms in the US before and after the GDPR.

Figure 5: Wedge Estimates



(a) Average Wedge by Industry

(b) Wedge Distribution

Notes: This figure presents our estimation results for the wedge induced by the GDPR (λ_i). Panel (a) presents the average estimated wedge for firms within each industry. Panel (b) presents the full distribution of estimated wedges. Solid lines denote the 95% confidence intervals, and standard errors are calculated using 100 bootstrap repetitions at the firm level.

This indicates that data and compute are more complementary than traditional inputs. We view these estimates as a contribution to the production function literature, as there is limited empirical evidence on how firms use data despite its growing importance.

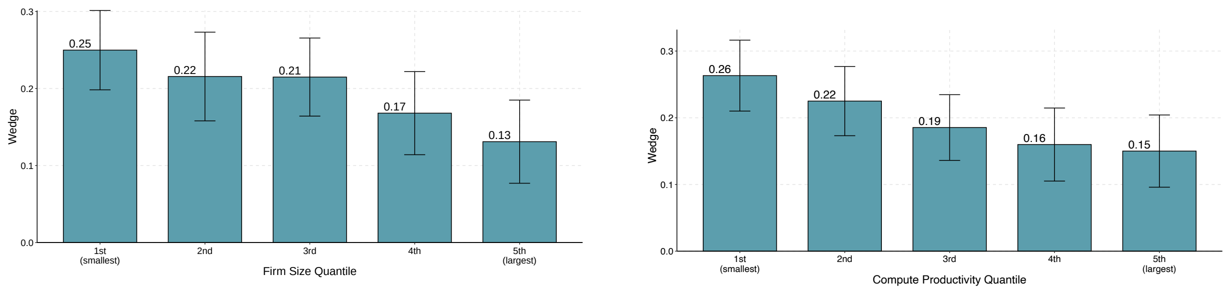
6.2 The Regulatory Wedge Induced by the GDPR

Next, we examine our estimates of the wedges introduced by the GDPR (λ_i). Panel (a) of Figure 5 displays the average wedge for EU firms across industries together with the 95% confidence intervals. The estimates are statistically significant and range from 17% to 24% across industries, implying that the GDPR raised the cost of data for firms. The wedge is the highest for software firms, likely due to their higher exposure to GDPR compliance requirements. These average estimates, however, hide substantial firm-level heterogeneity. As shown in Panel (b) of Figure 5, there is considerable heterogeneity in the wedge generated by the GDPR. For some firms, the wedge is close to zero, while for others, it can be as large as one.[40]

To better understand this heterogeneity and to study the determinants of these regulatory wedges, we look at how they correlate with two firm-level characteristics: (i) firm size, as measured by the number of employees, and (ii) compute-augmenting productivity, as

[40] Around 8% of our wedge estimates are negative, which we attribute to noise in the estimation.

Figure 6: Wedge Heterogeneity by Firm Size and Compute-Augmenting Productivity



(a) Average Wedge by Firm Size

(b) Average Wedge by ω_i^c

Notes: Figure presents our estimation results for the wedge induced by the GDPR (λ_i), averaging across firms within each of the given groups. Panel (a) shows these estimates across the five firm-size quintiles, while Panel (b) shows these estimates across the five compute-augmenting productivity (ω_i^c) quintiles computed using pre-GDPR estimates. Solid lines denote the 95% confidence intervals, and standard errors are calculated using 100 bootstrap repetitions at the firm level.

measured by ω_i^c estimates. The results are reported in Figure 6. Panel (a) shows the average wedge estimates across the five firm-size quintiles, where the quintiles are calculated within each industry. The results suggest that the distortionary effects of the GDPR are highest for the smallest firms (25%), with monotonically decreasing effects as the firm size gets bigger. This finding is consistent with other evidence on the effects of the GPPR in the literature (Campbell et al., 2015; Koski and Valmari, 2020; Goldberg et al., 2023) and may reflect the fact that larger firms have better resources to comply with the GDPR. In panel (b), we report the wedge distribution across quantiles of the compute-augmenting productivity distribution. As firms become more compute-intensive, the magnitude of the wedge decreases monotonically from 26% in the first quantile to 15% in the last quantile.

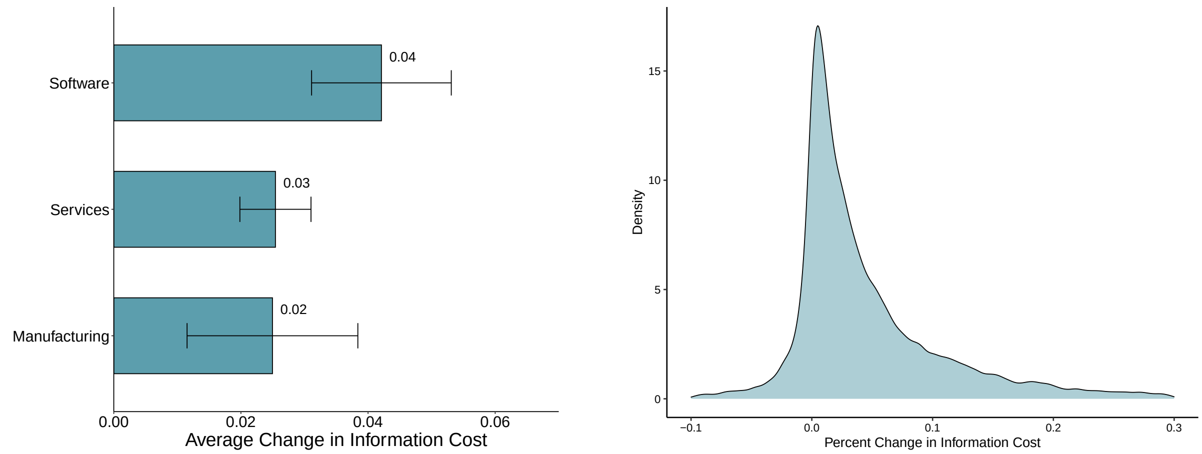
6.3 The Effect of the GDPR on the Cost of Information

How do the additional data costs resulting from the GDPR affect firms' production costs? To answer this question, we begin by deriving the effects of wedges on the "cost of information", the cost of producing a given level of information. This cost function can be derived from the CES production function as follows:

$$CI^*(I_{it}, p_{it}, \lambda_i) = I_{it} \left((\omega_{it}^c)^\sigma \left(p_{it}^c\right)^{1-\sigma} + \alpha^\sigma \left((1+\lambda_i)p_{it}^d\right)^{1-\sigma} \right)^{1/(1-\sigma)}, \tag{11}$$

with the derivation provided in Appendix E.4. This equation shows that the impact of λ_i on the information cost increases with data intensity (α), and decreases with the elasticity of substitution (σ).

Figure 7: Percent Increases in Information Costs



(a) Avg. Change in Info. Cost by Industry **(b)** Distribution of Changes in Information Cost

Notes: Figure presents the estimates of the percentage change in the cost of information induced by the GDPR calculated using Equation (11). Panel (a) presents the average estimated percentage increase in the cost of information for firms within each industry. Solid lines denote the 95% confidence intervals, and standard errors are calculated using 100 bootstrap repetitions at the firm level. Panel (b) presents the full distribution of the estimated percentage increase in the cost of information.

We use Equation (11) to estimate the percentage increase in the cost of information post-GDPR by considering two scenarios: (i) a case in which there was no wedge ($\lambda_i = 0$), so the cost of data is p_{it}^d, and (ii) the realized case in which the cost for firms included the costs of regulations: $(1 + \lambda_i)p_{it}^d$.[41] Using our estimates of model parameters, we calculate the ratio of (ii) to (i) for every firm-month at the estimated parameters (as prices and ω_{it}^c change month to month). This calculates the percentage change in information cost, which we further average to obtain firm-level measures.

The results for the percentage increases in information costs are reported in Figure 7. Panel (a) shows the average change by industry, plotting the mean along with standard errors. The average increase in the cost of information ranges from 2.5% for manufacturing to 4.2% in software, with significant firm-level heterogeneity reported in Panel (b).

Why does an average of 20% increase in the cost of data reported in the previous section lead to only a 2.5-4.2% increase in the information cost? To analyze this, we decompose

[41]Note that we can calculate the percentage increase in the cost for a given information level I, without taking into account the effects of changing I on information costs. The level of I would affect the unit cost of information when the information production function exhibits increasing or decreasing returns to scale. See Equation (19) in Appendix E.4 for more information.

the effects of λ_i on information cost as follows:

$$\frac{\mathrm{d}CI^*_{it}}{\mathrm{d}\lambda_i}\frac{\lambda_i}{CI^*_{it}} = \underbrace{s^d_{it}\,\lambda_i}_{\text{direct effect (+)}} + \underbrace{\left[s^d_{it}\left(\frac{\partial D^*_{it}}{\partial \lambda_i}\frac{\lambda_i}{D^*_{it}}\right) + \left(1 - s^d_{it}\right)\left(\frac{\partial C^*_{it}}{\partial \lambda_i}\frac{\lambda_i}{C^*_{it}}\right)\right]}_{\text{firm re-adjustment margin (-)}},$$

where s^d_{it} denotes the cost share of data in information production. In this decomposition, the first term—the direct effect—represents the increase in costs if firms do not re-optimize their data-compute input mix, while the second term—the firm re-adjustment margin—is the extent to which firms can mitigate the increase in costs by substituting data for compute while holding production fixed. Conceptually, if firms do not re-optimize their inputs, the increase in the cost of information would be determined by the expenditure share of data in information (s^d_{it}) multiplied by the wedge (hence the positive direct effect). However, firms' input re-optimization would reduce this effect depending on the elasticity of substitution (hence the negative re-adjustment margin).

Both channels explain why the cost of information increase is about a fifth of the average wedge. First, we find that the average direct effect is small at 3.9% because data expenditures account for only 19% of information production costs. The small expenditure share of data is an equilibrium outcome determined by both the data's role in the production function and its price relative to the compute. This observation—that firms allocate substantially more resources to computation than to data—provides an important insight into the role of data and computation in firm production.

Looking at the re-adjustment margin, we find that given the strong complementarity of data and compute, firms are limited in their ability to mitigate the increase in the information cost by substituting data for compute. Therefore, the average firm re-adjustment margin is only -0.2% (see Figure OA-5(b) for the distribution), contributing minimally to the overall effect of the GDPR on the cost of information.

To summarize, the small increase in the cost of information primarily comes from the small expenditure share of data in information production, with the re-optimization margin having little impact. Overall, this section highlights the importance of understanding the firm production with data to quantify the cost of privacy regulations.

6.4 The Effect of the GDPR on Firm Production Costs

Up until now, we have limited the scope of our analysis to the firm's production of information. In this subsection, we sacrifice some generality to analyze how changes in information costs translate into changes in production costs using simple back-of-the-

Table 7: Effects of the GDPR on Production Costs by Industry

	Software (1)	Services (2)	Manufacturing (3)
	Panel A. Key Parameter Values		
Increase in Information Costs (ΔCI_i)			
Mean increase	0.04	0.03	0.02
5^{th} - 95^{th} *percentile increase*	[0.03 - 0.05]	[0.02 - 0.04]	[0.01 - 0.03]
Elasticity of Substitution ($\bar{\sigma}$)			
Lashkari et al. (2023)	0.83	0.18	0.17
Information Expenditure Share (s_i^I)			
Median share	11.8%	5.0%	3.1%
Range of estimates	8.7% - 16.7%	2.9% - 5.0%	2.3% - 3.3%
	Panel B. Estimation Results		
Increase in Production Costs (ΔC_i)			
Mean increase	0.47%	0.15%	0.06%
Range of estimates	[0.26% - 0.82%]	[0.06% - 0.20%]	[0.02% - 0.10%]

Notes: Table presents estimates of Equation (13) calibrated with increases in the cost of information estimated in Section 6.3 and information expenditure shares estimated from Aberdeen and other industry surveys for each industry. The mean increase in production costs is calculated with the mean increase in information costs and the median information expenditure share. The range of estimates is calculated by combining the 5^{th} - 95^{th} percentile increases in information costs with the lower and upper range of information expenditure share estimates, respectively. Columns (1), (2), and (3) show estimates for software firms (SIC 7370-7377), services firms, and manufacturing firms, respectively. Appendix G provides more detail about the information expenditure share estimates, the point estimates of $\bar{\sigma}$ taken from Lashkari et al. (2024).

envelope calculations under additional assumptions.

We follow Lashkari et al. (2024) by using a nested CES production technology, where information I is combined with non-information inputs such as capital and labor, $M(L, K)$. We denote the production function by:

$$Y_i = \omega_i \big(\beta I_i^{\bar{\rho}} + (1-\beta) M_i^{\bar{\rho}}\big)^{1/\bar{\rho}}, \tag{12}$$

where ω_i denotes firm productivity, β denotes information intensity in production, and $\bar{\sigma} = 1/(1-\bar{\rho})$ represents the elasticity of substitution between information and non-information inputs. We drop the time subscript since we conduct this analysis cross-sectionally.

We show in Appendix G.1 that under some simplifying assumptions—that all inputs are flexible, firms are price takers in the input market, and that firms do not have market power—we can derive how information cost changes translate into production costs by using sufficient statistics. More explicitly, the expenditure share of information in total production cost (s_i^I) and elasticity of substitution between information and non-information inputs ($\bar{\sigma}$) are sufficient statistics for the effect of the GDPR on production costs:

$$\Delta C_i = \left((1 + \Delta CI_i)^{1-\bar{\sigma}} \cdot s_i^I + 1 - s_i^I\right)^{1/(1-\bar{\sigma})} - 1, \tag{13}$$

where ΔC_i denotes the percentage increase in variable production costs due to the percentage increase in the cost of information (ΔCI_i). Equation (13) reveals intuitive comparative statics: a given increase in (ΔCI_i) translates into larger increases in production costs for larger information shares (s_i^I) and lower elasticities ($\bar{\sigma}$).

Now, we turn towards estimating ΔC_i. We note that we previously calculated ΔCI_i at the firm level in Section 6.3. We will use its mean value for our baseline estimates in this section and its 5-95th percentile to establish bounds. The remaining parameters in Equation (13) are the elasticity of substitution between information and non-information inputs ($\bar{\sigma}$) and the information expenditure shares s_i^I. For the elasticity of substitution, we rely on the estimates by Lashkari et al. (2024), who estimated the elasticity of substitution between IT and non-IT input using firm-level data in different industries.[42] We follow this approach because the estimation of $\bar{\sigma}$ requires information on non-IT inputs, which we do not observe fully. The Lashkari et al. (2024) estimates, reported in Table 7, suggest that information and non-information inputs are complements in all industries.

For the information expenditure shares s_i^I, the estimates are difficult to calculate directly at the firm level, as most production datasets do not provide detailed information on IT-related inputs. Instead, we calculate expenditure shares at the industry level by using the Aberdeen dataset and various industry-level surveys, which we discuss in detail in Appendix G.2.[43] While we might expect each source to suffer from distinct drawbacks, we find that the sources generate remarkably consistent estimates across industries. Table 7 reports the median and the interquartile range of these estimates, whereas Table OA-10 provides the estimates from each source separately.

We present the estimated ranges for ΔC_i from Equation (13) in Panel B of Table 7. We estimate that production costs increase by 0.47% for software firms due to the GDPR. These increases are significantly larger than corresponding increases in the services and manufacturing industry, which we estimate as 0.15% and 0.06%, respectively. This difference is primarily driven by the larger information expenditure shares of software firms: the median expenditure share estimate for software firms is 11.8% compared to 5.0% for services and 3.1% for manufacturing firms. This difference is compounded by the fact that

[42] Lashkari et al. (2024) study France from 1995 - 2007. Although their setting predates ours, their comprehensive data on firm-level information technology investment and industry-level parameter estimates provide useful information on production functions with IT and non-IT inputs.

[43] While these sources only partially capture the information expenditure share and capture different samples of firms, we aim to provide a range of plausible values by combining estimates across surveys and years.

software firms also face the largest average wedges and resulting increases in the cost of information.[44]

To provide a sense of the quantitative magnitudes associated with our estimated increases in production costs, we multiply our estimates (ΔC_i) by the amount of GDP accounted for by each industry in the Euro Area in 2018.[45] This exercise implies an annual variable production cost increase for the software industry on the order of €3 billion. Furthermore, although service and manufacturing industries experienced smaller relative increases in production costs, the importance of these industries implies associated annual GDPR costs on the order of €11.8 and €1.2 billion, respectively.[46]

Although these calculations rely on strong assumptions, we view these results as informative in showing how the economic costs estimated from our production function translate into aggregate costs across different industries.

7 Conclusions

In this paper, we examine the impact of the GDPR on firm data input choices and their production costs. Comparing EU firms affected by the GDPR to similar firms in the US, we document that firms stored 26% less data and did 15% less computation two years after the GDPR, becoming less data-intensive. Our results contribute to the literature documenting the potential costs of the GDPR, complementing the existing literature by focusing on data inputs in firm production that have rarely been studied.

To map the observed shifts in input choices to changes in firms' production costs, we also propose a production function in which firms aggregate data and computation through a CES functional form. Estimates of this production function suggest that data and computation are strong complements in production. We then model the cost of the GDPR as a wedge between the marginal product of data and its price and find that the GDPR drove an average increase in the variable cost of data of 20%, with small firms experiencing more significant cost increases.

Using our estimates of production model parameters, we find that these increases in data costs translate into an average increase of only 3.7% in the variable costs of "infor-

[44]This exercise does not take into account the effects of economy-wide reallocation of production between firms (Oberfield and Raval, 2021) and other general equilibrium responses (Lashkari et al., 2024) which can be quantitatively important.

[45]Our estimates of the GDP accounted by each industry (and their share of the GDP) are €639 billion (5.53%), €7.84 trillion (67.86%), and €1.95 trillion (16.88%) for software, services, and manufacturing, respectively. We discuss how we attribute GDP to industries in greater detail in Appendix G.2.

[46]These numbers are in the same ballpark as some of the available estimates from surveys. For example, Ernst & Young estimated that in 2018, the largest 500 corporations in the world were on track to spend a total of $7.8 billion to comply with the GDPR (Bloomberg Businessweek, 2018).

mation." This relatively modest effect, despite an average 20% increase in data input costs, stems primarily from data's smaller expenditure share in firm production relative to computation. Finally, by assuming that the firm production takes a nested-CES form in information and other inputs, we estimate that these wedges imply a 0.06% increase in production costs for manufacturing firms and substantially larger increases around 0.47% for more data-intensive software firms.

Our results reinforce the importance of studying the impact of privacy regulation on firm production, and they emphasize the importance of considering "data usage" as a multi-dimensional object in a firm production with data and computation. We leave several important margins for future research, including studying the fixed costs of compliance and multi-national firms. We reiterate, however, that this paper is only a partial analysis of the welfare effects of the GDPR, as we are completely agnostic to the benefits that consumers derive from privacy protections. A full welfare analysis must incorporate these benefits into a single estimation framework.

References

Accenture (2018). Supercharging HR Data Management. Last accessed on 2023-01-05, https://www.accenture.com/t20180829t083931z__w__/hk-en/_acnmedia/pdf-85/accenture-supercharging-hr-financial-services.pdf.

Accenture (2022). To the multi-cloud and beyond. Last accessed on 2023-11-21, https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-180/accenture-to-the-multi-cloud-and-beyond.pdf.

Acemoglu, D. (2002). Directed Technical Change. *The Review of Economic Studies 69*(4), 781–809.

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification Properties of Recent Production Function Estimators. *Econometrica 83*(6), 2411–2451.

Acquisti, A., C. Taylor, and L. Wagman (2016). The Economics of Privacy. *Journal of Economic Literature 54*(2), 442–92.

Agrawal, A., J. Gans, and A. Goldfarb (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.

Agrawal, A., J. McHale, and A. Oettl (2019). Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, Volume I, Chapter 5, pp. 149–174. The University of Chicago Press.

Aridor, G., Y.-K. Che, and T. Salz (2023). The Effect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR. *RAND Journal of Economics 54*(4), 695–730.

Arrieta-Ibarra, I., L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl (2018). Should We Treat Data as Labor? Moving Beyond "Free". *AEA Papers and Proceedings 108*, 38–42.

Asker, J., A. Collard-Wexler, and J. De Loecker (2019). (Mis) allocation, Market power, and Global Oil Extraction. *American Economic Review 109*(4), 1568–1615.

Athey, S., C. Catalini, and C. Tucker (2017). The Digital Privacy Paradox: Small Money, Small Costs, Small Talk. *NBER Working Paper* (w23488).

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020). The Fall of the Labor Share and the Rise of Superstar Firms. *The Quarterly Journal of Economics 135*(2), 645–709.

Bartik, T. J. (1991). *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute.

Berger, D., K. Herkenhoff, and S. Mongey (2022). Labor Market Power. *American Economic Review 112*(4), 1147–1193.

Bertrand, M., C.-T. Hsieh, and N. Tsivanidis (2021). Contract Labor and Firm Growth in India. *NBER Working Paper* (w29151).

Bessen, J., S. M. Impink, L. Reichensperger, and R. Seamans (2022). The Role of Data for AI Startup Growth. *Research Policy 51*(5), 104513.

Bloom, N., R. Sadun, and J. V. Reenen (2012). Americans Do IT Better: US Multinationals and the Productivity Miracle. *American Economic Review 102*(1), 167–201.

Bloomberg Businessweek (2018). It'll Cost Billions for Companies to Comply With Europe's New Data Law. Last accesed on 2024-01-14, https://www.bloomberg.com/news/articles/2018-03-22/it-ll-cost-billions-for-companies-to-comply-with-europe-s-new-data-law?

Borusyak, K., P. Hull, and X. Jaravel (2022). Quasi-Experimental Shift-Share Research Designs. *The Review of Economic Studies 89*(1), 181–213.

Boyne, S. M. (2018). Data Protection in the United States. *The American Journal of Comparative Law 66*(1), 299–343.

Brand, J. M., M. Demirer, C. Finucane, and A. A. Kreps (2024). Firm Productivity and Learning in the Digital Economy: Evidence from Cloud Computing. *NBER Working Paper* (w32938).

Brynjolfsson, E. and L. M. Hitt (2003). Computing Productivity: Firm-Level Evidence. *Review of Economics and Statistics 85*(4), 793–808.

Byrne, D., C. Corrado, and D. E. Sichel (2018). The Rise of Cloud Computing: Minding Your P's, Q's and K's. *NBER Working Paper* (w25188).

Caballero, R. J., E. M. R. A. Engel, J. C. Haltiwanger, M. Woodford, and R. E. Hall (1995). Plant-Level Adjustment and Aggregate Investment Dynamics. *Brookings Papers on Economic Activity 1995*(2), 1–54.

Campbell, J., A. Goldfarb, and C. Tucker (2015). Privacy Regulation and Market Structure. *Journal of Economics & Management Strategy 24*(1), 47–73.

Canayaz, M., I. Kantorovitch, and R. Mihet (2022). Consumer Privacy and Value of Consumer Data. *Swiss Finance Institute Research Paper* (22-68).

Chan, M. (2023). How Substitutable are Labor and Intermediates? *Working Paper*.

Chander, A., M. Abraham, S. Chandy, Y. Fang, D. Park, and I. Yu (2021). Achieving Privacy: Costs of Compliance and Enforcement of Data Protection Regulation. *World Bank Policy Research Working Paper* (9594).

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2007). Business Cycle Accounting. *Econometrica 75*(3), 781–836.

Chen, L., Y. Huang, S. Ouyang, and W. Xiong (2021). The Data Privacy Paradox and Digital Demand. *NBER Working Paper* (w28854).

Chirinko, R. S. (2008). σ: The Long and Short of it. *Journal of Macroeconomics 30*(2), 671–686.

Choi, J. P., D.-S. Jeon, and B.-C. Kim (2019). Privacy and Personal Data Collection with Information Externalities. *Journal of Public Economics 173*, 113–124.

DataGrail (2020). The Cost of Continuous Compliance: Benchmarking the Ongoing Operational Impact of GDPR & CCPA. Last accessed on 2023-01-05, `https://www.datagrail.io/resources/reports/gdpr-ccpa-cost-report/`.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The Rise of Market Power and the Macroeconomic Implications. *The Quarterly Journal of Economics 135*(2), 561–644.

Demirer, M. (2022). Production Function Estimation with Factor-Augmenting Technology: An Application to Markups. *Working Paper*.

DeStefano, T., R. Kneller, and J. Timmis (2023). Cloud Computing and Firm Growth. *The Review of Economics and Statistics*, 1–47.

Dibble, S. (2019). *GDPR for Dummies*. John Wiley & Sons.

Doerr, S., L. Gambacorta, L. Guiso, and M. Sanchez del Villar (2023). Privacy Regulation and Fintech Lending. *BIS Working Papers* (1103).

Doraszelski, U. and J. Jaumandreu (2018). Measuring the Bias of Technological Change. *Journal of Political Economy 126*(3), 1027–1084.

Eeckhout, J. and L. Veldkamp (2022). Data and Markups: A Macro-Finance Perspective. *NBER Working Paper* (w30022).

European Commission (2008). Flash Eurobarometer 226: Data protection in the European Union: Data controllers' perceptions. Technical report, European Commission.

Ewens, M., R. Nanda, and M. Rhodes-Kropf (2018). Cost of Experimentation and the Evolution of Venture Capital. *Journal of Financial Economics 128*(3), 422–442.

Farboodi, M. and L. Veldkamp (2022). A Model of the Data Economy. *NBER Working Paper* (w28427).

Flexera (2023). State of the Cloud Report. Last accesed on 2023-06-19, `https://info.flexera.com/CM-REPORT-State-of-the-Cloud`.

Frey, C. B. and G. Presidente (2024). Privacy Regulation and Firm Performance: Estimating the GDPR Effect Globally. *Economic Inquiry 62*(3), 1075–1089.

Ganglmair, B., J. Kr"amer, and J. Gambato (2024). Regulatory Compliance with Limited Enforceability: Evidence from Privacy Policies. *ZEW-Centre for European Economic Research Discussion Paper* (24-012).

García-Dorado, J. L. and S. G. Rao (2015). Cost-aware Multi Data-Center Bulk Transfers in the Cloud from a Customer-Side Perspective. *IEEE Transactions on Cloud Computing 7*(1), 34–47.

GDPR.eu (2019). GDPR Small Business Survey. Last accessed on 2023-01-05, https://gdpr.eu/2019-small-business-survey/.

Gellert, R. (2018). Understanding the Notion of Risk in the General Data Protection Regulation. *Computer Law & Security Review 34*(2), 279–288.

Godinho de Matos, M. and I. Adjerid (2022). Consumer Consent and Firm Targeting after GDPR: The Case of a Large Telecom Provider. *Management Science 68*(5), 3330–3378.

Goldberg, S. G., G. A. Johnson, and S. K. Shriver (2023). Regulating Privacy Online: An Economic Evaluation of the GDPR. *American Economic Journal: Economic Policy 16*(1), 325–58.

Goldfarb, A. and C. Tucker (2012). Shifts in Privacy Concerns. *American Economic Review 102*(3), 349–53.

Goldfarb, A. and C. Tucker (2019). Digital Economics. *Journal of Economic Literature 57*(1), 3–43.

Goldfarb, A. and C. E. Tucker (2011). Privacy Regulation and Online Advertising. *Management Science 57*(1), 57–71.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020, August). Bartik Instruments: What, When, Why, and How. *American Economic Review 110*(8), 2586–2624.

Graetz, G. and G. Michaels (2018). Robots at Work. *Review of Economics and Statistics 100*(5), 753–768.

Greenleaf, G. (2022). Now 157 Countries: Twelve Data Privacy Laws in 2021/22. *SSRN Working Paper*.

Greenstein, S. M. and T. P. Fang (2020). *Where the Cloud Rests: The Economic Geography of Data Centers*. Harvard Business School.

Hicks, J. R. (1932). *The Theory of Wages*. Macmillan and Co Ltd., London.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and Manufacturing TFP in China and India. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Hughes, J. T. and A. Saverice-Rohan (2017). IAPP-EY Annual Privacy Governance Report 2017. Last accessed on 2013-06-19, https://iapp.org/media/pdf/resource_center/IAPP_EY_Governance_Report_2017.pdf.

Hughes, J. T. and A. Saverice-Rohan (2018). IAPP-EY Annual Privacy Governance Report 2018. Last accessed on 2023-01-05, https://iapp.org/media/pdf/resource_center/IAPP_EY_Governance_Report_2018.pdf.

Hughes, J. T. and A. Saverice-Rohan (2019). IAPP-EY Annual Privacy Governance Report 2019. Last accessed on 2013-06-19, https://iapp.org/media/pdf/resource_center/IAPP_EY_Governance_Report_2019.pdf.

Hustinx, P. (2013). EU Data Protection Law: The Review of Directive 95/46/EC and the Proposed General Data Protection Regulation. *University of Tartu. Data Protection Inspectorate, Tallinn*.

Ichihashi, S. (2020). Online Privacy and Information Disclosure by Consumers. *American Economic Review 110*(2), 569–95.

IT Governance Privacy Team (2017). *EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide - Second edition* (2 ed.). IT Governance Publishing.

Janßen, R., R. Kesler, M. Kummer, and J. Waldfogel (2021). GDPR and the Lost Generation of Innovative Apps. *NBER Working Paper* (w30028).

Jia, J., G. Z. Jin, and L. Wagman (2021). The Short-Run Effects of the General Data Protection Regulation on Technology Venture Investment. *Marketing Science 40*, 661–684.

Jin, W. (2022). Cloud Adoption and Firm Performance: Evidence from Labor Demand. *SSRN Working Paper*.

Jin, W. and K. McElheran (2017). Economies Before Scale: Survival and Performance of Young Plants in the Age of Cloud Computing. *Managment Science* (Forthcoming).

Johnson, G. (2022). Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond. *NBER Working Paper* (w30705).

Johnson, G., S. Shriver, and S. Goldberg (2023). Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR. *Management Science 69*(10), 5695–6415.

Jones, C. I. and C. Tonetti (2020, September). Nonrivalry and the Economics of Data. *American Economic Review 110*(9), 2819–2858.

Jones, M. L. and M. E. Kaminski (2020). An American's Guide to the GDPR. *Denver Law Review 98*(1), 93.

Kehrig, M. and N. Vincent (2021). The Micro-Level Anatomy of the Labor Share Decline. *The Quarterly Journal of Economics 136*(2), 1031–1087.

Kircher, T. and J. Foerderer (2020). Does EU-Consumer Privacy Harm Financing of US-App-Startups? Within-US Evidence of Cross-EU-Effects. In *Proceedings of the 42nd International Conference on Information Systems (ICIS), Austin, United States, December 12–15*, pp. 1–18.

Kirov, I. and J. Traina (2023). Labor Market Power and Technological Change in US Manufacturing. *Working Paper*.

Koski, H. and N. Valmari (2020). Short-Term Impacts of the GDPR on Firm Performance. *ETLA Working Papers* (77).

Krähmer, D. and R. Strausz (2023). Optimal Non-linear Pricing with Data-Sensitive Consumers. *American Economic Journal: Microeconomics 15*(2), 80–108.

Lashkari, D., A. Bauer, and J. Boussard (2024). Information Technology and Returns to Scale. *American Economic Review 114*(6), 1769–1815.

Lefrere, V., L. Warberg, C. Cheyre, V. Marotta, and A. Acquisti (2022). Does Privacy Regulation Harm Content Providers? A Longitudinal Analysis of the Impact of the GDPR. *SSRN Working Paper*.

Lin, T. and A. Strulov-Shlain (2023). Choice Architecture, Privacy Valuations, and Selection Bias in Consumer Data. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2023-58).

Loertscher, S. and L. M. Marx (2020). Digital Monopolies: Privacy Protection or Price Regulation? *International Journal of Industrial Organization 71*, 102623.

Lukic, K., K. M. Miller, and B. Skiera (2023). The Impact of the General Data Protection Regulation (GDPR) on the Amount of Online Tracking. *SSRN Working Paper*.

Mell, P., T. Grance, et al. (2011). The NIST Definition of Cloud Computing. Last accessed on 2013-06-19, https://csrc.nist.gov/publications/detail/sp/800-145/final.

Montes, R., W. Sand-Zantman, and T. Valletti (2019). The Value of Personal Information in Online Markets with Endogenous Privacy. *Management Science 65*(3), 1342–1362.

Morlacco, M. (2020). Market Power in Input Markets: Theory and Evidence from French Manufacturing. *Working Paper*.

Oberfield, E. and D. Raval (2021). Micro Data and Macro Technology. *Econometrica 89*(2), 703–732.

OECD (2020). *National Accounts of OECD Countries, Volume 2020 Issue 1*.

O'Kane, P. (2017). *GDPR-Fix it Fast: Apply GDPR to Your Company in 10 Simple Steps*. Brentham House Publishing Company Ltd.

Olley, G. S. and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica 64*(6), 1263–1297.

Peter, A. and C. Ruane (2023). The Aggregate Importance of Intermediate Input Substitutability. *NBER Working Paper* (w31233).

Peters, M. (2020). Heterogeneous Markups, Growth, and Endogenous Misallocation. *Econometrica 88*(5), 2037–2073.

Peukert, C., S. Bechtold, M. Batikas, and T. Kretschmer (2022). Regulatory Spillovers and Data Governance: Evidence from the GDPR. *Marketing Science 41*, 746–768.

Ponemon Institute (2017). The True Cost of Compliance with Data Protection Regulations. Last accessed on 2023-06-19, https://static.fortra.com/globalscape/pdfs/guides/gs-true-cost-of-compliance-data-protection-regulations-gd.pdf.

Ponemon Institute (2018). The Race to GDPR. Last accessed on 2023-08-17, https://iapp.org/media/pdf/resource_center/Ponemon_race-to-gdpr.pdf.

Ponemon Institute (2019). Keeping Pace in the GDPR Race: A Global View of GDPR Progress. Last accessed on 2023-06-19, https://www.privacysecurityacademy.com/wp-content/uploads/2019/06/Keeping-Pace-in-the-GDPR-Race.pdf.

Raval, D. R. (2019). The Micro Elasticity of Substitution and Non-Neutral Technology. *The RAND Journal of Economics 50*(1), 147–167.

Restuccia, D. and R. Rogerson (2008). Policy Distortions and Aggregate Productivity with Heterogeneous Establishments. *Review of Economic Dynamics 11*(4), 707–720.

Schmitt, J., K. M. Miller, and B. Skiera (2022). The Impact of Privacy Laws on Online User Behavior. *HEC Paris Research Paper* (MKG-2021-1437).

Syverson, C. (2011). What Determines Productivity? *Journal of Economic Literature 49*(2), 326–365.

Tuzel, S. and M. B. Zhang (2021). Economic Stimulus at the Expense of Routine-Task Jobs. *The Journal of Finance 76*(6), 3347–3399.

Veldkamp, L. and C. Chung (2023). Data and the Aggregate Economy. *Journal of Economic Literature 62*(2), 458–84.

Voigt, P. and A. Von dem Bussche (2017). The EU General Data Protection Regulation (GDPR). *10*(3152676), 10–5555. Publisher: Springer.

Zhao, Y., P. Yildirim, and P. K. Chintagunta (2021). Privacy Regulations and Online Search Friction: Evidence from GDPR. *SSRN Working Paper* (3903599).

Zhuo, R., B. Huffaker, kc claffy, and S. Greenstein (2021). The Impact of the General Data Protection Regulation on Internet Interconnection. *Telecommunications Policy 45*(2), 102083.

Zolas, N., Z. Kroff, E. Brynjolfsson, K. McElheran, D. N. Beede, C. Buffington, N. Goldschlag, L. Foster, and E. Dinlersoz (2021). Advanced Technologies Adoption and Use by US Firms: Evidence from the Annual Business Survey. *National Bureau of Economic Research, No. 28290*.

Data, Privacy Laws & Firm Production: Evidence from GDPR

Mert Demirer, Diego Jiménez-Hernández, Dean Li and Sida Peng

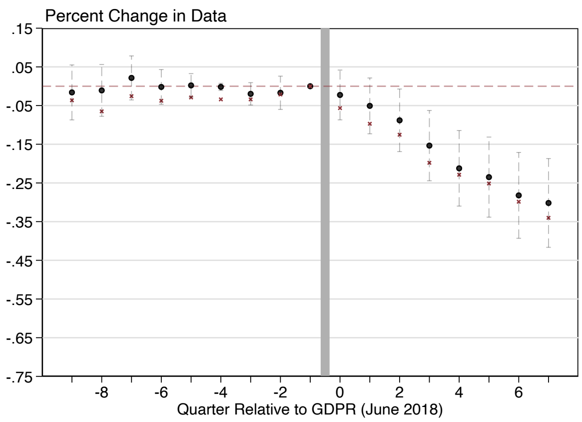
Appendix - For Online Publication

Contents

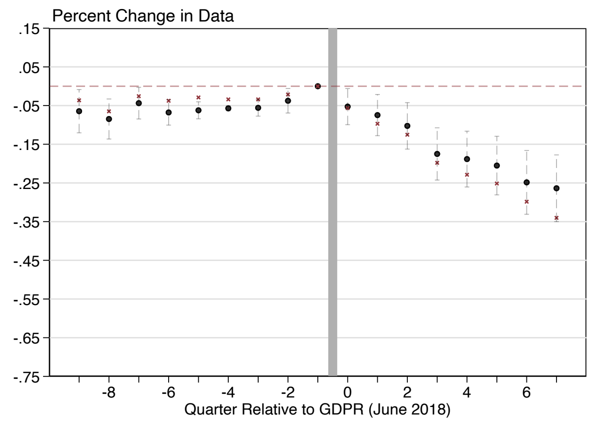
A Additional Exhibits

Figure OA-1: Event Study Estimates of the Effect of the GDPR on Cloud Inputs
(Effects on Storage by Industry)

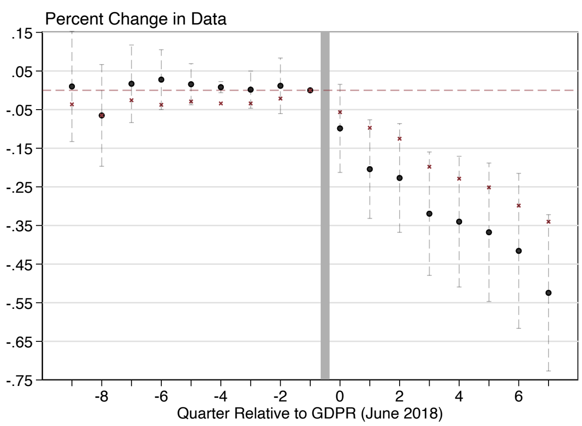
(a) Software



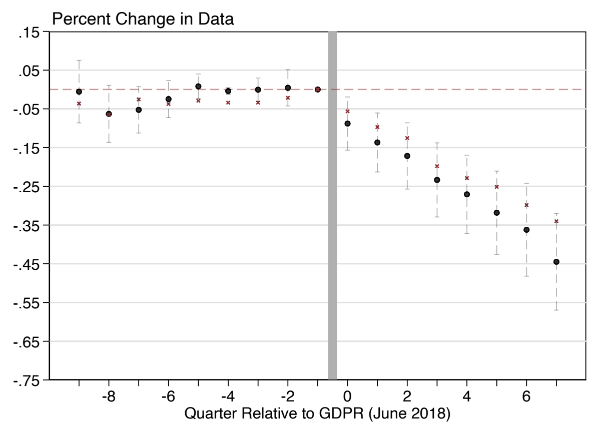
(b) Services



(c) Manufacturing



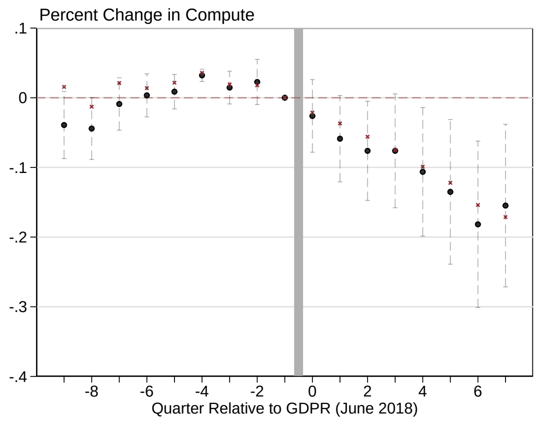
(d) Other



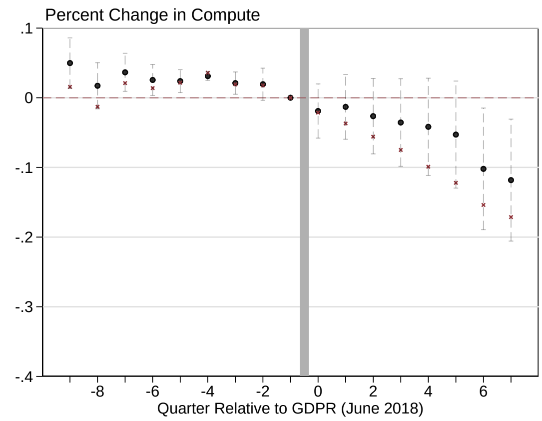
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator, when the outcome is log storage. The coefficient in the quarter before the GDPR's implementation is normalized to zero. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. Results are broken down by industry, and red dots show the main estimates from the paper. The full definition of industries and the corresponding observation numbers are available in Table 4.

Figure OA-2: Event Study Estimates of the Effect of the GDPR on Cloud Inputs (Effects on Compute by Industry)

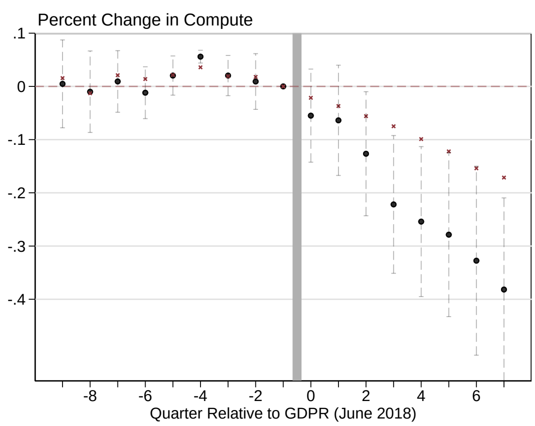
(a) Software



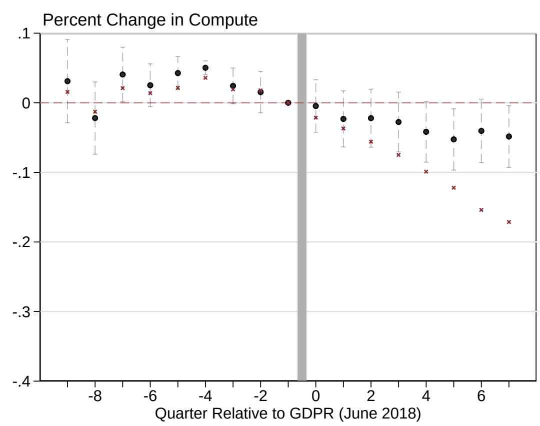
(b) Services



(c) Manufacturing

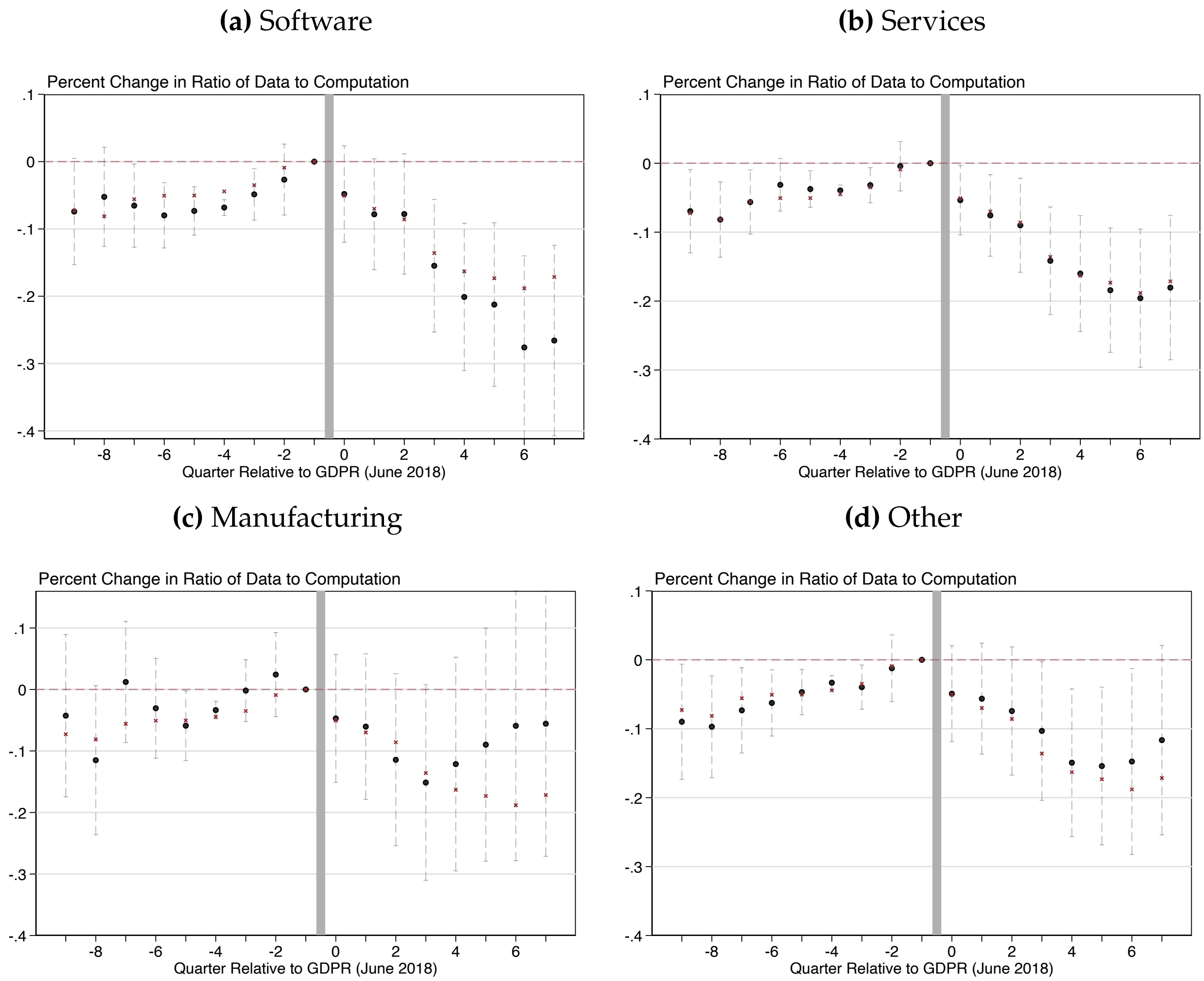


(d) Other



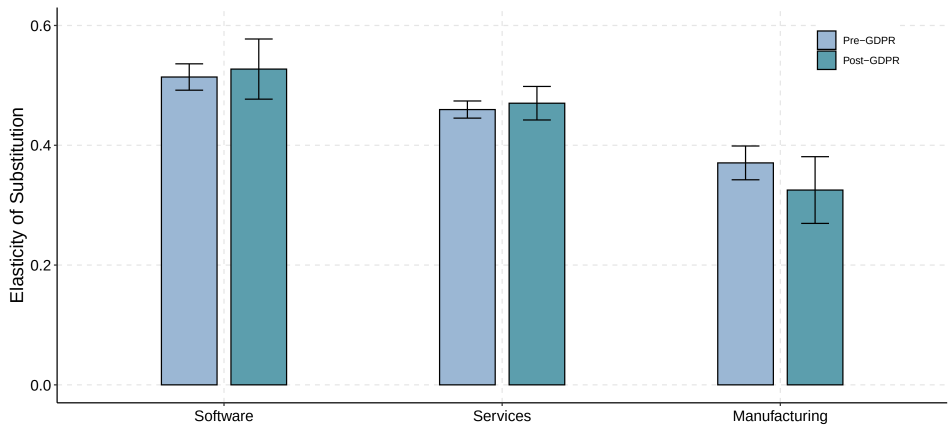
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator, when the outcome is log computation. The coefficient in the quarter before the GDPR's implementation is normalized to zero. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. Results are broken down by industry, and red dots show the main estimates from the paper. The full definition of industries and the corresponding observation numbers are available in Table 4.

Figure OA-3: Event Study Estimates of the Effect of the GDPR on Cloud Inputs (Effects on Data Intensity by Industry)



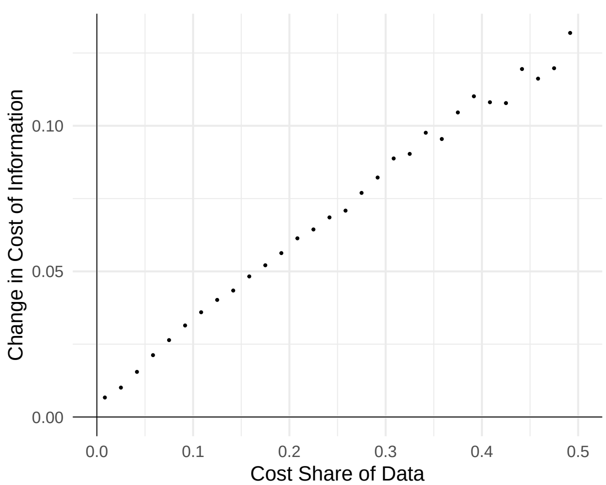
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator, when the outcome is log data intensity. The coefficient in the quarter before the GDPR's implementation is normalized to zero. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. Results are broken down by industry, and red dots show the main estimates from the paper. The full definition of industries and the corresponding observation numbers are available in Table 4.

Figure OA-4: Elasticity of Substitution Between Data and Compute for US Firms

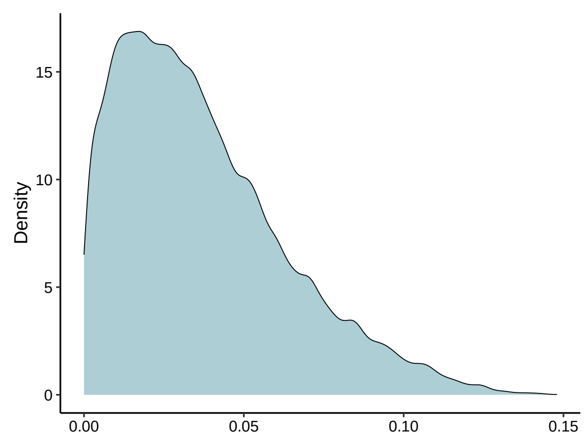


Notes: This table presents our estimation results of the elasticity of substitution between data and compute (σ) across industries. We present separate estimates for the pre- and post-GDPR (σ_1 and σ_2, respectively). Standard errors are calculated using 100 bootstrap repetitions.

Figure OA-5: Additional Results on Information Cost



(a) Avg. Change in Info. Cost by Data Share



(b) Firm Re-Adjustment Margin

Notes: Figure presents our additional estimation results for the change in the cost of information induced by the GDPR. Panel (a) presents the average estimated increase in the cost of information by the pre-GDPR level of the total expenditures in data. Panel (b) shows our estimates of the "firm re-adjustment" contribution to the total change in the cost of information, computed firm by firm as the difference between the increase in the cost of information and the first-order approximation given by the data expenditure share times the wedge.

B The Impact of the GDPR on Firms

B.1 GDPR Summary

In this section, we present a more detailed description of the GDPR. In particular, we focus on the main changes that firms must implement to comply with the GDPR. This section is compiled from information presented in IT Governance Privacy Team (2017), Dibble (2019), Voigt and Von dem Bussche (2017), O'Kane (2017) and original GDPR legal text.

Definition of Controller and Processor (Article 4). A controller is defined as an entity that determines the purposes and means of processing personal data. A processor, on the other hand, is defined as an entity that processes personal data on behalf of a controller. Under the GDPR, a processor is not considered a third party, so the controller can involve a processor at its discretion and does not need a legal basis to do so. If a processor is chosen, it must be suitable and provide sufficient guarantees to implement appropriate technical and organizational measures that meet GDPR requirements and protect data subjects' rights. Both parties must enter into a written contract or other legal agreement to bind the processor to the necessary conditions.

Records of Processing Activities (Article 30). Controllers and processors must create records of their processing activities that include details on the purposes of processing, the categories of data being processed, and descriptions of the technical and organizational security measures in place. There are exceptions to record-keeping requirements for organizations with fewer than 250 employees unless the processing it carries out is likely to result in a risk to the rights and freedoms of data subjects, the processing is not occasional, or the processing includes special categories of data.

Designation of a Data Protection Officer (Article 37). The GDPR requires data controllers and processors to designate a Data Protection Officer (DPO) in the following cases: (i) the processing is carried out by a public authority or body, except for courts acting in their judicial capacity; (ii) the core activities of the controller or the processor involve regular and systematic monitoring of data subjects on a large scale; (iii) the core activities of the controller or the processor consist of processing on a large scale of special categories of data listed in Article 9 and Article 10.

Preparing a Data Protection Impact Assessment (Article 35). If an intended processing activity, especially one involving new technologies, is likely to result in a high risk to the rights and freedoms of data subjects, then firms must conduct a Data Protection Impact Assessment (PIA) to identify and implement appropriate measures to mitigate privacy risks. The PIA should be conducted at the start of a project so that all stakeholders are aware of any potential privacy risks. The PIA should include the following components: (i) a systematic description of the purposes and planned processing operations, including the controller's legitimate interests (if applicable); (ii) an assessment of the necessity and proportionality of the processing in relation to the purpose; (iii) an assessment of the risks to the rights and freedoms of the data subjects; and (iv) the measures planned to address these risks.

Technical and Organizational Measures for Data Security (Article 32). The controllers must put technical and organizational measures in place to protect personal data. They should implement appropriate data protection policies that are proportionate to their processing activities with a risk-based approach. The GDPR does not specify a specific set of security controls that firms must implement but rather encourages data controllers and processors to implement "appropriate" controls based on risk.

Data Subject Rights (Article 14-21). Under the GDPR, individuals have extensive rights when their personal data is collected by data controllers. These rights include requesting data erasure, data transfer, and data access. If a request is made by a data subject, the firm must respond to the request without undue delay and generally within one month of receiving the request. As a result, firms may need to proactively fulfill a number of obligations so that they can quickly provide information about their processing, erase personal data, provide or transfer specific data, or correct incomplete personal data.

Data Breach Notification (Article 33). Under the GDPR, controllers have a general duty to report personal data breaches to supervisory authorities within 72 hours of becoming aware of them. When a personal data breach is likely to result in a high risk to the rights and freedoms of natural persons, the controller must notify the affected data subjects without undue delay.

Penalties and Increased Liability Risk (Article 83). The GDPR makes it easier for data subjects to bring civil claims against data controllers and processors. The data subject does not need to have suffered financial loss or material damage (e.g., loss or destruction of goods or property) to bring a claim. They can also claim for non-material damage, such as distress or hurt feelings. The GDPR sets out two levels of administrative fines. The higher level of fines can be up to €20 million or 4% of the total global annual turnover of the preceding financial year, whichever is higher. This level applies to infringements of certain fundamental principles, such as individuals' basic rights and freedoms. The lower level of fines can be up to €10 million or 2% of the total global annual turnover of the preceding financial year, whichever is higher. This level applies to other types of infringements.

B.2 The Compliance Cost of the GDPR

Compliance with the GDPR is likely to create significant costs for firms. Some of these costs are one-time fixed costs that are associated with actions required for initial compliance with the GDPR, while others are ongoing variable costs required for continuous compliance. In this section, we present evidence highlighting the impact of the GDPR on firm costs collected from various firm surveys. See Chander et al. (2021) for an overview of the costs of compliance associated with data privacy laws for businesses.

Although there are no official statistics available on the overall costs of the GDPR, surveys provide information on the cost of compliance with GDPR regulations. The estimates range from an average of \$3 million (Hughes and Saverice-Rohan, 2018) and \$5.47 (Ponemon Institute, 2017) to \$13.2 million (Ponemon Institute, 2019) depending on the composition surveyed firms. Importantly, the responses to these surveys indicate that these costs do not consist solely of one-time costs, and firms expect to incur these costs repeatedly (Ponemon Institute, 2019). Studies that provide a breakdown of these costs indicate that a high percentage of the costs (between one-fifth and one-half) are the labor costs of privacy compliance personnel. Depending on the study, technology accounts for 12 to 17% of total GDPR cost, and outside consultants and lawyers account for another 19 to 24% (Ponemon Institute, 2019; Hughes and Saverice-Rohan, 2019).

B.2.1 Fixed and Sunk Costs

Operational Changes for Data Security and Processing The GDPR potentially requires many operational changes from firms, such as implementing data protection management systems. These changes involve sunk and fixed costs. The cost component associated with operational changes can be quite large, independent of the quantity of data a firm has or uses. This is because firms must develop and implement technical and organizational measures to comply with potential consumer requests and other reporting requirements for data breaches. Other components of fixed costs include data mapping, writing privacy notices, and training employees on GDPR compliance.

Data Protection Officer The GDPR requires a data protection officer (DPO) for some firms, depending on their data processing activity. Even though DPO is a primarily fixed cost, it can also be seen as a variable cost since the number of employed DPOs can increase with firm size and data. A survey by IAPP with 370 respondents suggests that 18% of firms have appointed multiple DPO (Hughes and Saverice-Rohan, 2017), indicating that DPO could be a variable cost for large firms.

B.2.2 Variable Costs

Some of the costs associated with GDPR compliance are variable and scale with the size of the organization and the amount of data it possesses. According to a survey conducted by DataGrail, 88% of firms spend over $1 million, and 12% spend more than $10 million annually to maintain GDPR compliance (DataGrail, 2020). The heterogeneity in continuous compliance costs suggests that some costs are variable and change with firm size and amount of stored data. Below, we provide some examples of variable GDPR compliance costs.

Handling Customer Requests Under the GDPR, consumers have the right to have their data erased, transferred, or even made available for their downloading. The costs of handling these requests are likely to be variable, as companies with more data are more likely to receive requests. Survey evidence supports this conclusion. According to (DataGrail, 2020), 58% of companies receive more than 11 customer requests per month, and 28% receive more than 100 requests. More than half of companies have at least 26 employees managing these requests. Moreover, only 1% of companies report fully automating these requests, with 64% handling them entirely manually.

Recording Data Processing Activities An important aspect of the GDPR is creating a plan for new projects that involve data collection and processing. For example, if a firm needs to implement a new machine learning algorithm with new variables, it must do a detailed analysis for risk assessment, cost-benefit analysis, and necessary safeguards to prevent potential future issues. This constitutes a significant project-specific cost that might affect the cost-benefit trade-off for implementing new data collection projects. Therefore, some projects that involve data might not be implemented due to this additional cost.

Improved Data Security Keeping data in a more secure environment can also increase the variable cost of data, especially for cloud computing users. Cloud providers offer different tiers of security for their storage services, with higher levels of security typically corresponding to higher costs. Purchasing these additional storage services due to the GDPR would increase the marginal cost of storing data for firms.

Liabilities The maximum penalties under the GDPR include fines of up to €20 million or 4% of the company's global annual revenue, whichever is greater. However, the actual penalty amount is determined by the nature and severity of the violation and is likely to be increasing with the amount of data stored by the firm. Moreover, one can imagine that the probability of a cyberattack could increase with the amount of data. Another related variable cost is cybersecurity insurance. Of the 1,263 organizations surveyed in Ponemon

Institute (2019), 31% of respondents purchased cyber-risks insurance. Of those insured, 43% had insurance coverage for GDPR fines and penalties.

B.3 Publicly Available GDPR Fine Data

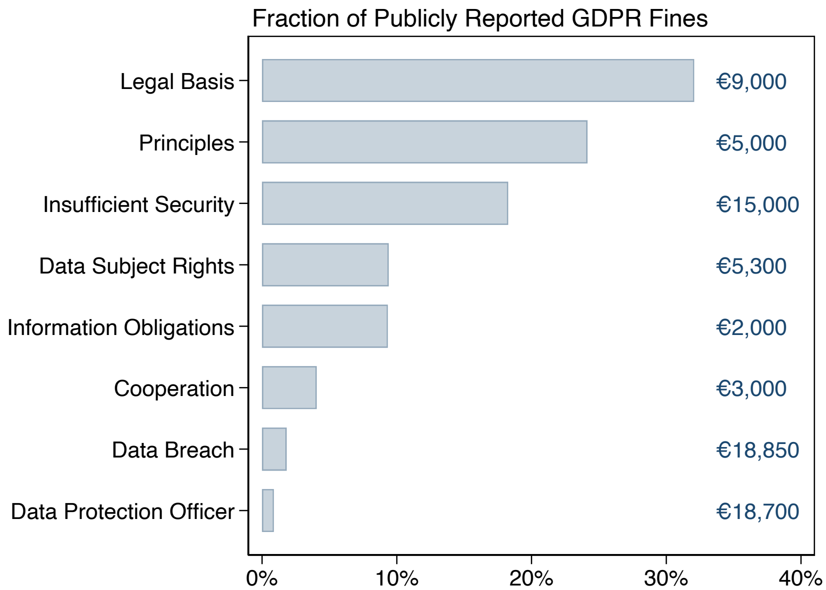
Our primary source of publicly available fine data is a database maintained by CMS Legal Services, a large international law firm that operates in over 40 countries. This data provides an overview of the public fines and penalties that data protection authorities have imposed under the GDPR. Although not all fines are made public, the data on public fines is quite rich, containing the fine amount, the entity being fined, the country of the fine, and the GDPR articles under which the fine was leveled.[47] The database contains more than €3 billion in fines levied in the five years after the implementation of the GDPR. Furthermore, there are primary and secondary sources associated with each of the fines in the database.

For each fine, we scrape the fine amount, the entity it was levied on, the date, and why the fine was levied. In Figure 1 in the paper, we show the distribution of fine sizes, highlighting that there is considerable variation in the size of the fines. There is also substantial variation in the specific reasons that fines were levied, and these reasons fall into eight categories: (a) insufficient legal basis for data processing, (b)insufficient involvement of data protection officer, (c) insufficient technical and organizational measures to ensure information security, (d) insufficient fulfillment of information obligations, (e) non-compliance with general data processing principles, (f) insufficient fulfillment of data subjects rights, (g) insufficient cooperation with the supervisory authority, and (h)insufficient fulfillment of data breach notification obligations. For brevity, we label these as “legal basis", “data protection officer", “data security", “information obligations", “data principles", “data subject rights", “non-cooperation", and “data breach notifications" respectively.

In Figure OA-6, we show the share of fines that were levied under each reason and the median fine size conditional on the reason. Perhaps unsurprisingly, data security concerns result in the largest types of fines. The median fines for insufficient information security and insufficient notification of data breaches are €15,000 and €18,850, respectively, while the median fines for non-cooperation and insufficient fulfillment of information obligations are €3,000 and €2,000 respectively. Overall, the distribution of the reasons given for the publicly available GDPR fines suggests that fines may be levied against firms for various reasons.

[47] We scraped this data in May 2023 through `https://www.enforcementtracker.com/`.

Figure OA-6: Publicly Reported GDPR Fines



Notes: Figure presents the distribution of reasons given for GDPR fines, using the publicly reported fine data described in Appendix Section B.3. Fine reasons are derived from the GDPR Article quoted in the fine, and these reasons are broken out into eight categories by CMS Law. We drop the 1.5% of fines that have no quoted GDPR article. Appendix Section B.3 describes these categories in further detail. The median fine size by reason is provided in blue text on the right side of the figure.

C Data Appendix

C.1 Cloud Computing Details

This section details how firms perform computation and storage in cloud computing, which is the main focus of our paper.

C.1.1 Computation

Firms requiring cloud computation typically opt for virtual machines (VMs). VMs are a type of cloud computing "compute" product that allows users to create and manage servers virtually instead of maintaining their own physical hardware.[48] These VMs run on virtualized infrastructure provided by a cloud computing provider and can access software and computing resources. These machines are typically fully customizable and controlled by the user. Cloud computing VMs can be configured in various ways. Some of the features of VMs that can be customized include memory, storage, networking options, CPU, operating system, and the data center's location. Cloud computing providers offer hundreds of different configurations, and the user chooses the exact configuration when requesting a VM.

In our paper, we use the number of CPU cores in a virtual machine as the key measure of computation outcome because this is the key vertical VM characteristic that determines compute capacity. However, we note that this approach does not consider heterogeneity in other characteristics, such as memory, networking capacity, and VM manufacturer/series.

The unit of observation is "core-hours", which refers to the amount of computing time a VM uses over a given period. The number of core-hours is calculated by multiplying the number of CPU cores by the number of hours the VM runs. For example, if a user runs a VM with 4 CPU cores for 10 hours, the total compute use would be 4 x 10 = 40 core hours. Cloud providers typically use core-hours as the primary measure of VM usage for billing purposes.

C.1.2 Storage

Cloud providers offer a wide range of storage products that can be used for various purposes, including storing and managing data, backing up and recovering data, and archiving data for long-term retention. These products can be categorized into two types: disk storage and database storage. Disk storage provides physical hardware where firms

[48] There are other "compute" products—such as containers and serverless computing—that were also available during our sample period, but they were not extensively used.

can store a wide variety of data, including operating system files, applications, documents, and multimedia files. Disk storage can include different physical configurations, such as Hard Disk Drives (HDDs) and Solid-State Drives (SSDs), as well as Storage Area Networks. Disk storage can also differ based on other characteristics, such as upload and download speed. Databases, on the other hand, are collections of structured data that are hosted and managed in a cloud computing environment by a cloud provider. The differentiation of databases refers to the various types of databases available and their specific features and characteristics, such as MySQL, NoSQL, Oracle, and PostgreSQL.

Firms typically use storage in one of two ways. First, when a firm creates a VM on a cloud provider's infrastructure, it can choose the amount of disk storage it needs and specify the required performance characteristics. They would use this disk storage when computing on that virtual machine. Second, firms might request either disk storage or databases to store and manage application data, and this storage might be used for supporting real-time applications and services or as archiving storage.

Our unit of observation for storage is storage capacity measured in gigabytes (GB). It represents a direct measure of how much data firms store, although it does not measure how storage products may be vertically or horizontally differentiated. An important example of storage differentiation is upload and download speed.

C.2 Sample Selection and Cleaning

In this section, we discuss our sample construction in greater detail. We define a firm as a unique internal identifier for whom we are able to observe industry classification and location information. Using this definition, we are able to capture approximately 90% of storage and 95% of computation in our entire sample.

Next, we clean the data by removing outlying observations. To tag a firm as an outlier, we require that we observe the firm's usage in the months immediately preceding and following a given month. We define outliers as large and sudden temporary spikes or temporary dips. These are months where a firm's usage is either twenty times more or less usage than the same firm's usage in the months immediately preceding and following the month. We also filter these by minimum size change to ensure that we are not spuriously removing small firms with more volatile usage. This cleaning removes less than 0.1 percent of observations. We also worked with internal employees to conduct some minor cleaning to remove a small fraction of firms whose observations are affected by the introduction and phaseout of older service models for our provider.

We then construct our sample by conditioning it on continuous firm observation for one full year, exactly two years before the GDPR. Although the resulting sample of firms is smaller, conditioning on the continuously observed firms does not significantly change the share of data that we observe. In fact, these continuously observed firms are responsible for about 90 percent of storage and computation before the GDPR. We present summary statistics on these sets of firms below in Table OA-1. While for confidentiality, we cannot provide direct comparisons between the number of firms before and after this conditioning, the mean storage and compute are given relative to a baseline normalization of 1,000 mean units of storage for our baseline sample in Table 2.

Table OA-1: Summary Statistics: Before Conditioning on Observation Period

Industry	Share of Firms	Share Compute	Share Storage	Mean Storage	Mean Compute	Share EU
Software	18.0	20.6	16.6	341	331	58.6
Services	47.1	34.5	38.6	408	296	38.2
Manufacturing	7.7	11.4	10.2	593	518	55.5
Other	27.2	33.6	34.6	651	479	49.7
All	100	100	100	468	345	46.3

Notes: Table presents summary statistics from our matched sample of firms. A description of the sample's construction can be found in Section 3.1, and a more detailed description of the sample construction can be found in Appendix C. This sample presents firms in Cases 1 and 4, as described in Table 1. For confidentiality purposes, we do not report the total number of firms. We also normalize the units of mean storage and mean computation such that everything is presented relative to a mean of 1,000 mean storage units in our baseline sample (Table 2).

C.3 Aberdeen Sample

Aberdeen is a market research firm that gathers data from various sources on firms' hardware and software investments. Every year, they survey a sample of senior IT executives about their software and hardware usage and extrapolate this information to non-surveyed firms. Additionally, they conduct large-scale data collection efforts, such as web scraping job postings and purchasing customer lists from vendors to identify software choices. Our understanding is that information on technology adoption comes only from the latter source. This data also includes sales, the number of employees, industry, and a DUNS number, which are sourced from Duns & Bradstreet. Our sample of Aberdeen data covers the period from 2015 to 2021 at the annual level. The data from Aberdeen has been previously used to study digitization and technology adoption (Graetz and Michaels, 2018; Tuzel and Zhang, 2021).

We use Aberdeen to measure the market shares of cloud providers in the EU and the US. Aberdeen provides information at two levels: the site level and the enterprise level.

A site refers to a physical location, while an enterprise corresponds to a firm (which may have multiple sites). The data includes unique site and enterprise IDs and a crosswalk that links the two. On average, the dataset covers more than 2 million sites, and the technology adoption information is reported at the site level. We aggregate this site-level information to the enterprise level by assuming that if at least one site of an enterprise uses a technology from a given provider, the enterprise uses the technology from that provider.

C.3.1 Match Procedure Between Aberdeen and Cloud Data

Aberdeen's data contains valuable information, such as revenue and employment, that we use to study the heterogeneity of our results and to illustrate how firms use the cloud. However, there is no single identifier we can use to match the anonymous cloud provider's data to Aberdeen, so we must resort to 'non-exact' procedures (also known as fuzzy matching) to link these two datasets. In both the cloud provider's and Aberdeen's data, we observe names, DUNS numbers (partial coverage in the cloud data), websites (URL), and partial address information, including postal codes, city, state, and country of the given firms. Additionally, we observe both the subsidiary name and the parent company's name in the Aberdeen data, which provides us with two potential strings to match each of our observations in our cloud data. Below, we provide details on the matching algorithm.

We use the Jaro-Winkler (JW) distance to match names, which considers the number of transpositions and the number of matching characters between two strings. Intuitively, strings with more characters in common and requiring fewer transpositions for one string to be contained within the other have lower distances. For the same number of character matches and transpositions, the JW distance is smaller for strings that match the first characters of the strings.[49]

For each firm in the cloud computing dataset, we find the "closest" match in the Aberdeen dataset (either by using the parents or the subsidiaries' names). We sequentially match using the following criteria and say that two firms are a match if both:

1. Share the same DUNS number, or
2. Share the same website, or
3. Are in the same postal code, and the name distance is less than 0.1, or
4. Are in the same city, and the name distance is less than 0.08, or
5. Are in the same stat,e and the name distance is less than 0.07, or

[49] In terms of the implementation, we use the Firm Merge Project (available at https://github.com/microsoft/firm-merge-project) to implement the JW distance in finite time.

6. Are in the same count,ry and the name distance is less than 0.065, or
7. Are in the same region (e.g., EU), and the name distance is less than 0.045.

Suppose a firm in the cloud computing data has multiple matches in the Aberdeen data. In that case, we hierarchize based on the same order as we list our criteria above.[50] Note that we also allow for "looser" string matching when the geographic region in which we search for a given firm is smaller. These cutoffs were chosen by visually inspecting the data and balancing the false-positive and false-negative matches.

With this procedure, we are able to match close to 60% of firms in our baseline sample to Aberdeen firms. We use this matched sample to study the heterogeneity of our result based on the firm's employment size. The change of firm employment over time is not as reliable at Aberdeen as the employment information does not change for a significant number of firms over time. For this reason, we use the employment information in 2018 to define firm size.

C.3.2 Aberdeen Cross-check with Internal Data

Even though Aberdeen was widely used to measure IT spending in the 2000s, the data has undergone changes in recent years, broadening its focus from hardware spending to software adoption. While hardware expenditure predominantly relied on surveys, the information on technology adoption at a larger scale mainly relies on web scraping, publicly available information, and extrapolation. This raises the question of how reliable the Aberdeen data is for technology adoption information. We find ourselves in a unique position to offer a partial answer to this question because we possess internal data from one of the largest cloud providers and can cross-check Aberdeen data for this provider.

To implement this, we utilize the matched Aberdeen-internal data sample to investigate whether Aberdeen accurately reports the adoption of our cloud computing provider. In particular, we examine the true positive and false negative rates: (i) the proportion of actual users of our cloud product that are correctly labeled and (ii) the proportion of users who do not use our cloud product that is correctly labeled. We find that the true positive rate is 64 percent, increasing with firm size, and the true negative rate is 66 percent, decreasing with firm size. This result suggests that while the Aberdeen data is not 100% accurate, it still provides a valuable signal about cloud adoption.

[50] For example, for a firm in the cloud computing data that we match by criteria (1) and (3) to different firms in the Aberdeen data, we only keep the match in criteria (1), given that DUNS numbers are designed as unique firm identifiers.

D Robustness Checks

This Appendix goes through the most critical threats to identification. We study substitution to other providers in Appendix D.1. We then investigate whether differential price changes (between the EU and the US) may be driving our results in Appendix D.2. We study firms with and without website usage (to measure the extent to which cookie collection drives our results) in Appendix D.3. Finally, we show that our results are robust to alternative choices of empirical strategies, sample selection procedures, and extensive margin/attrition in Appendix D.4.

D.1 Substitution to Other Providers

This section documents that substitution (to other cloud providers or to in-house IT services) is unlikely to drive our results. We provide a battery of exercises, each of which shows that substitution is unlikely to generate the patterns we observe in the data.

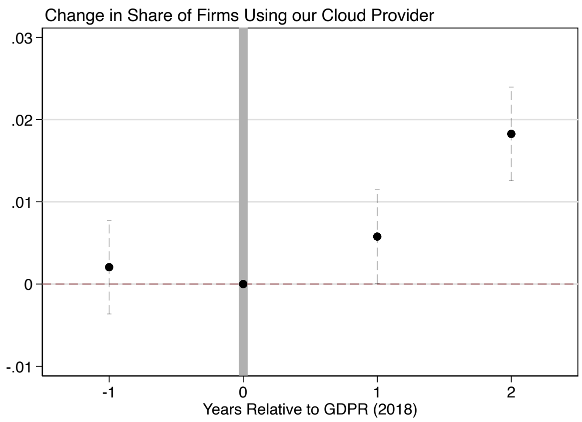
Substitution to Other Cloud Providers "Multi-cloud" usage—where firms get cloud services from multiple cloud computing providers—-is common among firms. Industry surveys suggest that 70 percent of cloud users are multi-cloud. Multi-cloud usage could be a potential issue because we observe usage from only one cloud computing provider, leading to incomplete data on cloud usage. If the GDPR changed the relative attractiveness between our cloud computing provider and other providers, perhaps in terms of how easily they accommodated GDPR regulations, then there could have been a differential change in our provider's market share in Europe and the US around the GDPR. This would pose an identification challenge for us.

In particular, we might attribute a decline in cloud storage and computing to firms simply switching their cloud usage to other providers. We note, however, that firms that conduct both storage and computing are likely to do both with the same provider because data cannot be stored with one provider but processed with another. For example, there are essentially no observations where a firm uses cloud computing with our provider without using cloud storage. Thus, our data intensity results should be less affected by any changes in the relative attractiveness of cloud providers.

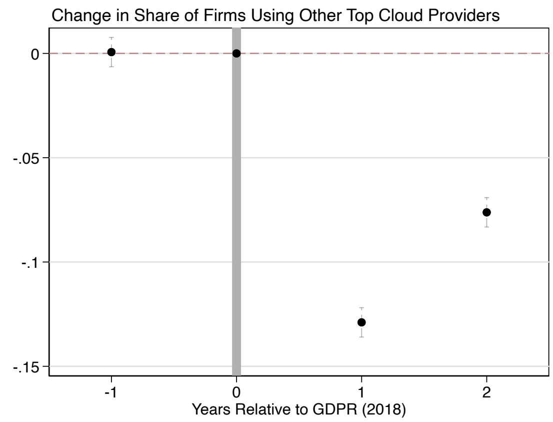
We attempt to address the identification challenge to our storage and computing results with three additional exercises. First, we bring an external dataset, Aberdeen, that provides information on firms' technology adoption and which vendors they get cloud services from. Using this dataset, we look at our provider's share of firms that receive services from each of the top cloud providers in Europe and US before and after GDPR and plot them in Figure OA-7. We find that the share of firms that are using our cloud provider has moderately increased over time, while the share of firms using the other cloud providers has decreased. Thus, we do not expect the relative attractiveness of the cloud provider that we observe to have decreased after GDPR.

Figure OA-7: Change in Share of Firms Using Cloud Providers in the EU vs the US

(a) Our Cloud Provider



(b) Other Cloud Providers



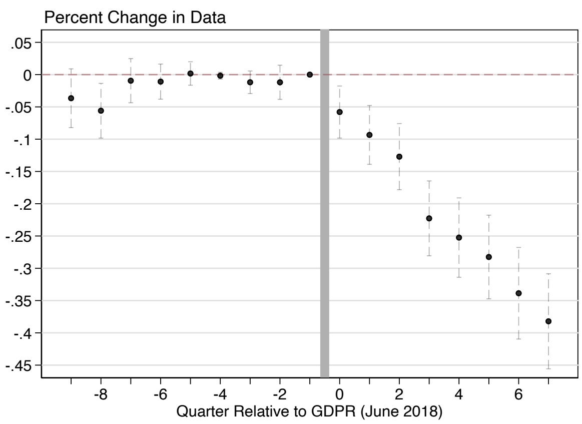
Notes: Figure presents estimates of the difference in the share of firms who use different cloud providers in the EU vs the US. The data source is Aberdeen (formerly known as Harte Hanks). The dependent variable on the left panel is equal to one if a firm uses the cloud provider that we study in this paper. The dependent variable in the right panel is equal to one if a firm uses any of the other cloud providers. The coefficients plot the difference in the share of firms who use the given cloud provider in the EU minus the share of firms using the same provider in the US, normalizing to the differences in 2018.

Second, we identify single cloud firms using Aberdeen again and estimate our empirical specification using only these firms. In Appendix C.3.2, we assess the reliability of Aberdeen data to identify these single-cloud firms and show that Aberdeen data provides useful information to detect single-cloud firms. Table OA-2 and Figure OA-8 present our estimates using this sample, which is quite similar to our baseline estimates across all outcomes. As discussed in the paper's main body, it is unlikely that the declines we observe are simply driven by substitution in usage to other providers.

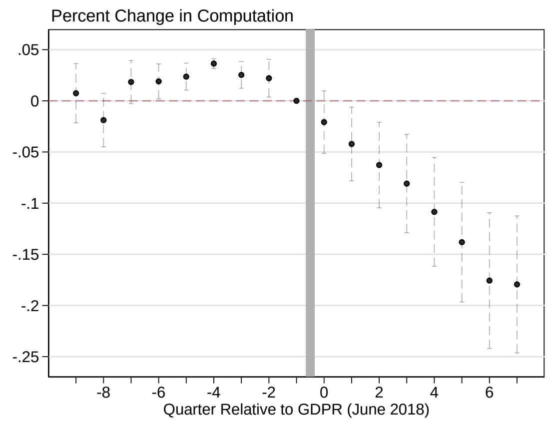
Finally, as discussed in Appendix B.1, the GDPR is likely to make multi-cloud usage more difficult. Thus, switching between cloud providers is more likely to occur on the *extensive* margin rather than the *intensive* margin. Thus, any cloud usage declines in a sample of firms that continuously use our provider are unlikely to be driven by switching between cloud providers. Table OA-3 presents estimates from a balanced panel of firms, where positive cloud computing usage is observed two years before and after the GDPR. These estimated coefficients for the short-run and long-run effects of the GDPR are quite similar to our baseline estimates. In particular, they are consistent with our findings

Figure OA-8: Event Study Estimates of the Effect of the GDPR on Cloud Inputs (Excluding Multi-Cloud Firms)

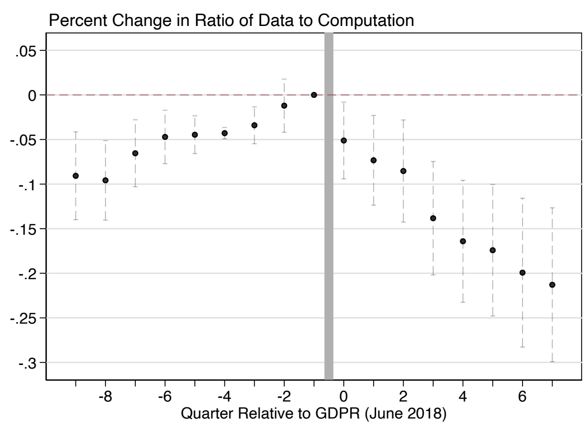
(a) Data



(b) Compute



(c) Data Intensity



Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. The outcome in each subpanel is denoted by the subpanel title. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. Sample sizes are presented in Table OA-2. The sample is composed of firms that do not use multiple cloud computing providers.

Table OA-2: Short- and Long-Run Effects of the GDPR
(Excluding Multi-Cloud Firms)

	Data (1)	Compute (2)	Data Intensity (3)
Short-Run Effect	-0.128	-0.085	-0.061
	(0.020)	(0.019)	(0.023)
Long-Run Effect	-0.258	-0.170	-0.121
	(0.027)	(0.028)	(0.034)
Observations	944,982	530,123	328,973
US Firms	13,166	7,891	4,152
EU Firms	14,112	7,415	4,832

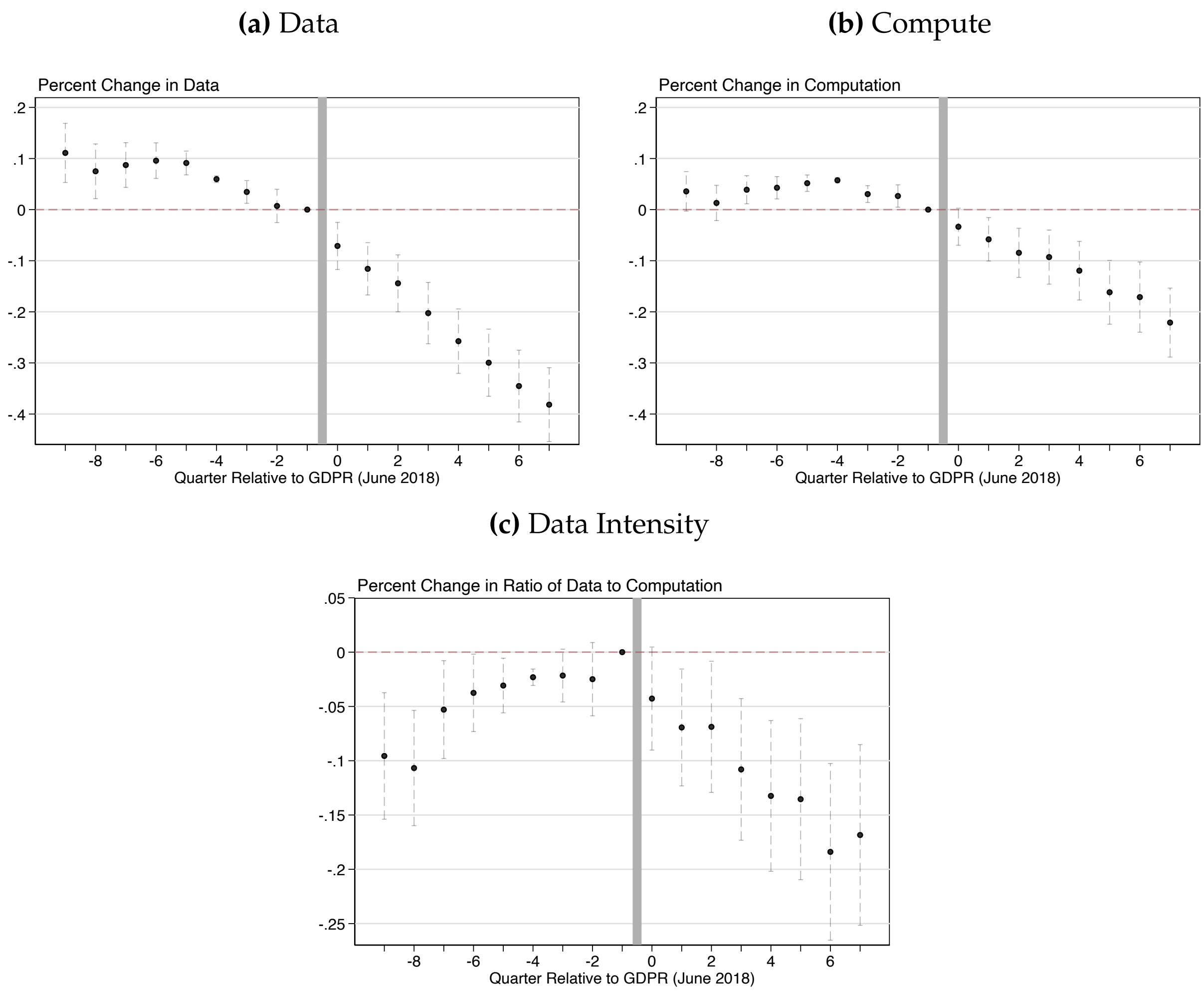
Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. Column (1) estimates the effect on storage. Column (2) estimates the effect on computation. Column (3) presents estimates of the data intensity. The sample excludes multi-cloud firms as described in Appendix D. Industries are defined as the ten divisions classified by SIC codes, with the addition of a "software" division, which we carve out of the services division and define through SIC codes 7370 - 7377. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we define "size decile" as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the firm level.

of a large decrease in both compute and storage alongside a decrease in data intensity. Thus, the results from our balanced panel in Table OA-3 and Figure OA-9 suggest that the declines in computation and storage we observe are not driven by switching between providers.

Substitution to On-Premises IT Next, we consider that firms might use both on-premises IT and cloud computing. To the extent that we cannot observe on-premises IT usage, declines in cloud computing may reflect re-allocations towards on-premises IT rather than true declines in computing. While increasing cloud computing adoption rates suggest that this margin may not play an important role, we consider the possibility that after the GDPR was enacted, European firms might have changed allocation between cloud and on-premises IT differently from US firms.

This would invalidate our identification arguments for the effects of compute and storage, although it would not necessarily affect the results on data intensity. To provide a robustness check for this, we focus on start-ups, which are unlikely to be using on-premises IT. These are young software firms that are less likely to switch toward on-premise IT than more established firms due to the sizable upfront costs. In Table OA-4 and Figure OA-10, we actually find larger effects for these firms rather than smaller effects. This suggests that the observed declines in computing and storage are unlikely to be driven by substitution

Figure OA-9: Event Study Estimates of the Effects of the GDPR on Cloud Inputs (Balanced Panel Estimates)



Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. The outcome in each subpanel is denoted by the subpanel title. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. Sample sizes are presented in Table OA-2. The sample is a balanced panel, and details can be found in Appendix Section D.

Table OA-3: Short- and Long-Run Effects of the GDPR
(Balanced Panel Estimates)

	Data (1)	Compute (2)	Data Intensity (3)
Short-Run Effect	-0.221	-0.115	-0.046
	(0.024)	(0.020)	(0.027)
Long-Run Effect	-0.373	-0.205	-0.104
	(0.030)	(0.029)	(0.037)
Observations	608,562	363,793	227,022
US Firms	7,588	5,126	2,872
EU Firms	7,953	4,112	2,849

Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. Column (1) estimates the effect on storage. Column (2) estimates the effect on computation. Column (3) presents estimates of the data intensity. The sample is a balanced panel, which is constructed as described in Appendix D. Industries are defined as the ten divisions classified by SIC codes, with the addition of a "software" division, which we carve out of the services division and define through SIC codes 7370 - 7377. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we define "size decile" as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the firm level.

to on-premises IT.

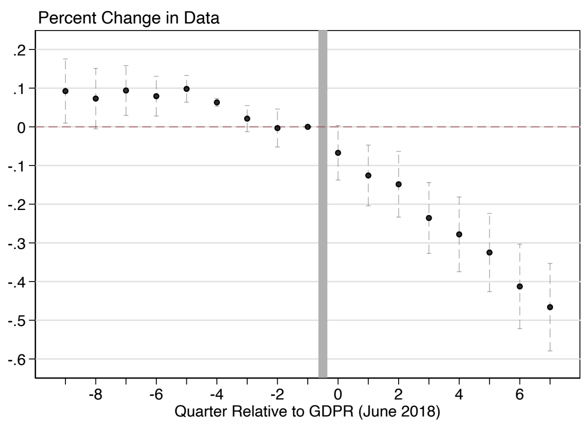
Table OA-4: Short- and Long-Run Effects of the GDPR
(Start-Up Firms)

	Data (1)	Compute (2)	Data Intensity (3)
Short-Run Effect	-0.241	-0.100	-0.069
	(0.036)	(0.027)	(0.034)
Long-Run Effect	-0.424	-0.202	-0.165
	(0.047)	(0.040)	(0.049)
Observations	311,128	267,066	157,616
US Firms	4,550	4,101	2,190
EU Firms	3,819	3,179	1,974

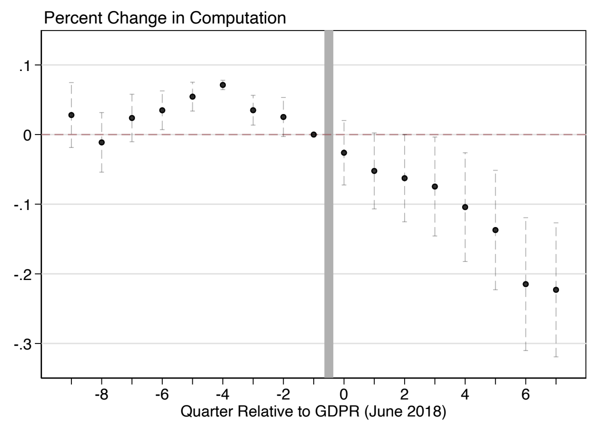
Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. Column (1) estimates the effect on storage. Column (2) estimates the effect on computation. Column (3) presents estimates of the data intensity. The sample is composed of start-up firms, classified according to a definition internal to the cloud provider. Industries are defined as the ten divisions classified by SIC codes, with the addition of a "software" division, which we carve out of the services division and define through SIC codes 7370 - 7377. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we define "size decile" as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the firm level.

Figure OA-10: Event Study Estimates of the Effects of the GDPR on Cloud Inputs (Start-Up Firms)

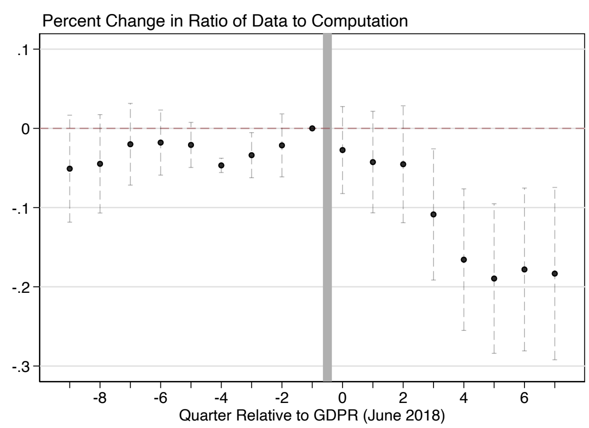
(a) Data



(b) Compute



(c) Data Intensity



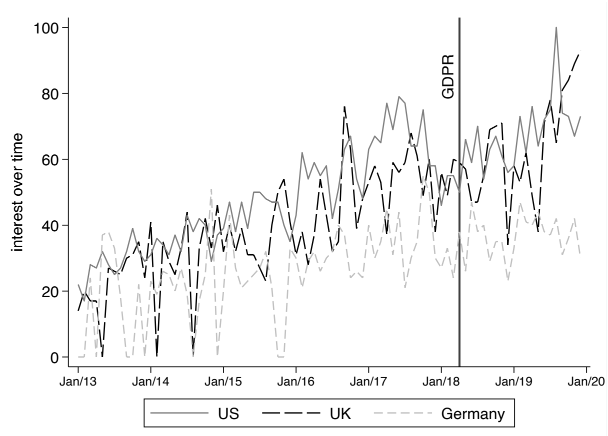
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. The outcome in each subpanel is denoted by the subpanel title. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. Sample sizes are presented in Table OA-4. The sample is composed of start-up firms, where start-ups are labeled according to a definition internal to the cloud provider.

Hybrid Cloud To further consider whether differential hybrid cloud usage after GDPR could explain our results, we explore its importance across regions and compare it with the importance of cloud computing using Google Trends data. Google Trends compares the volume of search topics for a given term (e.g., "hybrid cloud") after anonymizing and standardizing data. We downloaded data for "hybrid cloud" and "cloud computing" in the United States, the United Kingdom, and Germany. Results are in Figure OA-11.

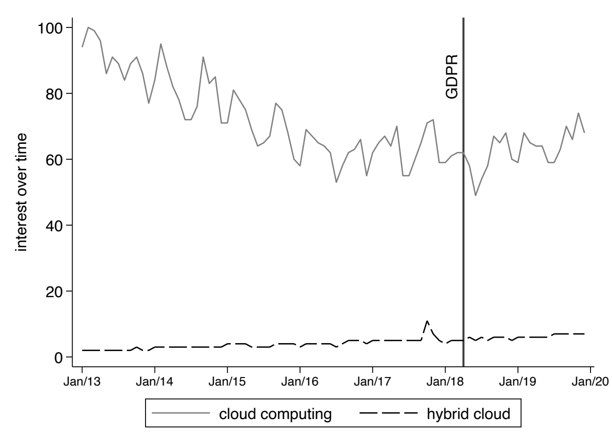
First, we compare the relative importance of hybrid cloud in Europe and in the US in Figure OA-11(a). If differential take-up of hybrid cloud in the EU were to explain our results, then one would expect hybrid cloud searches to increase post-GDPR. We do not find evidence of this. Rather, relative interest in hybrid cloud computing in the EU, if anything, declines after the GDPR. Furthermore, although we focus on the United Kingdom and Germany in the EU due to language differences, results are similar if we include searches from Italy, Spain, or France (both in English and in their own language). Second, Figure OA-11(b) compares interest in hybrid cloud and cloud computing worldwide from 2013 to 2021. Note that cloud computing is about 8 - 12 times more important as a term than hybrid cloud both before (March 2018) and after (December 2020) the GDPR. Second,

Figure OA-11: Google Trends Data on Cloud Computing and Hybrid Cloud

(a) Hybrid Cloud in US vs. in Europe



(b) Cloud Computing vs. Hybrid Cloud



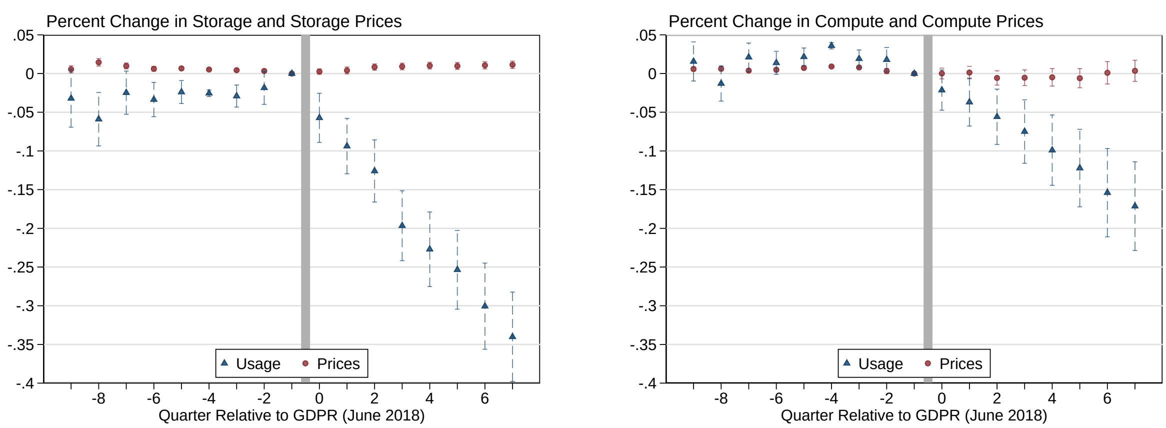
Notes: This figure compares Google Trends data for cloud computing and the hybrid cloud. Google Trends normalizes to 100 the topic-month with the most amount of searches. For example, a value of 50 on a given topic means that the topic is half as popular. Panel (a) plots the relative importance of the term "hybrid cloud" across the United States, the United Kingdom, and Germany. Panel (b) plots the relative importance of the terms "cloud computing" and "hybrid cloud" worldwide.

D.2 Price Changes

One natural channel through which the GDPR may have affected firms is through price changes in cloud computing. This would suggest our results might capture pricing responses by cloud providers rather than the GDPR's direct impact on firms. For example, if cloud computing providers increase their prices in the European Union relative to the United States, this could confound our estimates. While conversations with internal employees suggest that there were no explicit pricing responses to the passage of the GDPR, we also examine the data for evidence of any differential pricing trends between the EU and the US, either in listed or paid prices. Figure OA-12 presents our results when we estimate our event study specification using paid prices as the outcome. We find no evidence of significant differential price changes.

Figure OA-12: Event Study Estimates of the Effect of the GDPR on Cloud Inputs (Effects on Paid Prices)

(a) Storage Prices　　**(b)** Compute Prices



Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. The outcome in each subpanel is denoted by the subpanel title. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. The dependent variables shown in blue are our baseline estimates. The dependent variable shown in red is the paid price for each product.

D.3 Websites and Cookie Collection

One of the most salient aspects of the GDPR is the requirement that firms receive consent for the collection of data. This is particularly important in the case of websites and cookies: post-GDPR, websites that need to collect personal information must get explicit consent. As studied by Aridor et al. (2023), there may also be selection in terms of which consumers choose to opt out of data collection and how valuable the remaining data is.

We aim to study whether our main effects are driven by the GDPR's effect on websites and how important the selection channel might be for our sample. To examine whether or not web usage is driving our effects, we turn towards Table OA-5, where we proxy for active website use through the usage of cloud-based web services. These are services provided by our cloud provider that firms use to host their websites.

Re-estimating our empirical specification using firms with and without websites, we indeed find that firms using web services seem to have been more affected by the GDPR regulations: the effects on storage and computing are twice as large as those for non-active website users. However, the results remain statistically significant for non-active website users, and we additionally find that the adjustments in data intensity are similar. These results suggest that our effects are not solely driven by exposure to the GDPR's web-based cookie consent requirements. Similarly, restricting our sample to firms with no listed websites (regardless of whether that website is hosted within our cloud provider) provides qualitatively similar results. Results for the latter are available upon request.

D.4 Additional Robustness Exercises

Alternative Empirical Specifications The analyses in Section 4 are robust to several alternative specifications, including running our specification at the monthly level, the exclusion of various fixed effects, and alternative log-like transformation specification choices. Table OA-6 presents our event study results when the time periods are defined at the monthly level rather than at the quarterly level. In our main specification, we estimate coefficients and fixed effects at the quarterly level to preserve data confidentiality and increase the precision of our estimates. We find that our estimated coefficients are stable when we allow time trends to vary flexibly at the monthly level. The magnitudes of the estimated declines in storage, declines in computation, and decreases in data intensity are all quite similar to our baseline results.

We also consider the robustness of our analysis to the exclusion of our fixed effects. Our baseline specification allows for time trends to vary flexibly by industry and pre-GDPR size deciles. In the paper's Table 3, we consider alternative fixed effect specifications, including allowing time trends to vary only by industry and pre-GDPR size deciles and not allowing them to vary at all. We continue to observe the same features of our baseline results, including large long-run declines in storage and compute and moderate decreases in data intensity.

Finally, we consider alternative log-like transformations. Our baseline specification

Table OA-5: Short- and Long-Run Effects of the GDPR
(Heterogeneous Effects by Usage of Cloud-Based Web Services)

	Baseline (1)	Web Users (2)	Non-Web Users (3)
Panel A. Dependent variable: Log of Data			
Short-Run Effect	-0.129	-0.242	-0.080
	(0.018)	(0.020)	(0.010)
Long-Run Effect	-0.257	-0.421	-0.174
	(0.024)	(0.024)	(0.015)
Observations	1,143,149	255,057	888,092
US Firms	16,409	3,632	12,777
EU Firms	16,281	3,166	13,115
Panel B. Dependent variable: Log of Compute			
Short-Run Effect	-0.078	-0.124	-0.026
	(0.016)	(0.011)	(0.010)
Long-Run Effect	-0.154	-0.241	-0.060
	(0.024)	(0.018)	(0.019)
Observations	672,942	343,286	329,656
US Firms	10,294	5,243	5,051
EU Firms	8,927	4,297	4,630
Panel C. Dependent variable: Log of Data Intensity			
Short-Run Effect	-0.072	-0.066	-0.084
	(0.020)	(0.013)	(0.013)
Long-Run Effect	-0.131	-0.118	-0.112
	(0.029)	(0.023)	(0.024)
Observations	418,804	198,352	220,452
US Firms	5,487	2,714	2,773
EU Firms	5,872	2,608	3,264

Notes: Table presents estimates of equation (2) of δ_1 and δ_2, splitting our sample separately into firms that were observed using cloud-based web services with our provider between 24 and 13 months before the GDPR and those which were not. For comparison, Column (1) presents our baseline estimates across the full sample. Standard errors are clustered at the firm level.

Table OA-6: Short- and Long-Run Effects of the GDPR
(Monthly Specification)

	Data (1)	Compute (2)	Data Intensity (3)
Short-Run Effect	-0.141	-0.085	-0.079
	(0.018)	(0.017)	(0.021)
Long-Run Effect	-0.291	-0.174	-0.136
	(0.026)	(0.027)	(0.033)
Observations	1,143,149	672942	418,803
US Firms	16,409	10,294	5,487
EU Firms	16,281	8,927	5,872

Notes: Table presents estimates of equation (2) of δ_1 and δ_2, but where we allow our time trends to vary at the monthly level rather than the quarterly-level. Industries are defined as the ten divisions classified by SIC codes, with the addition of a "software" division, which we carve out of the services division and define through SIC codes 7370 - 7377. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we define "size decile" as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the firm level.

uses $\log(x)$. In Table OA-7 below, we consider using *asinh* and $\log(x+1)$. We find essentially no difference between these transformations, suggesting that our results are not sensitive to the behavior of our outcome transformations around zero.

Table OA-7: Short- and Long-Run Effects of the GDPR
(Alternative Transformations)

	Baseline (1)	*Asinh* (2)	*Log(x + 1)* (3)
Storage:			
Short-Run Effect	-0.129	-0.129	-0.126
	(0.018)	(0.018)	(0.019)
Long-Run Effect	-0.257	-0.257	-0.253
	(0.024)	(0.025)	(0.026)
Compute:			
Short-Run Effect	-0.078	-0.077	-0.076
	(0.016)	(0.016)	(0.016)
Long-Run Effect	-0.154	-0.153	-0.153
	(0.024)	(0.024)	(0.025)

Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. Column (1) shows our baseline specification with the natural logarithm of x. Column (2) transforms outcomes using the inverse hyperbolic sine. Column (3) transforms outcomes by taking the logarithm (base 10) of $x + 1$.

Alternative Sample Definitions We also discuss the robustness of our analyses in Section 4 to alternative sample definitions. In particular, we show that our estimated coefficients are relatively stable when estimated when conditioning on a different window of pre-GPDR usage, and when using a larger and more inclusive definition of “firms" where we don’t require any internal or external industry or operating information.

First, we consider alternative windows of pre-GDPR usage. In our baseline sample, we use firms for whom we observe cloud usage continuously for a whole year exactly two years before the GDPR. Table OA-8 presents estimates from the samples constructed by instead conditioning on continuous observation one year before the GDPR (column 2) and both years before the GDPR (column 3).

Table OA-8: Short- and Long-Run Effects of the GDPR (Alternative Pre-GDPR Usage Windows)

	(1)	(2)	(3)
Data:			
Short-Run Effect	-0.129	-0.101	-0.144
	(0.018)	(0.029)	(0.024)
Long-Run Effect	-0.257	-0.283	-0.299
	(0.024)	(0.039)	(0.034)
Compute:			
Short-Run Effect	-0.078	-0.078	-0.083
	(0.016)	(0.021)	(0.021)
Long-Run Effect	-0.154	-0.178	-0.178
	(0.024)	(0.033)	(0.033)
Data Intensity:			
Short-Run Effect	-0.072	-0.066	-0.063
	(0.020)	(0.023)	(0.023)
Long-Run Effect	-0.131	-0.128	-0.121
	(0.029)	(0.035)	(0.035)
Usage Observed During Year:			
Two Years Before GDPR	✓		✓
One Year Before GDPR		✓	✓

Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. Column (1) shows our baseline specification. Column (2) conditions on observing firms for the year before GDPR (instead of two years before). Column (3) restricts the sample to firms continuously observed for the full two years before GDPR. Industries are defined as the ten divisions classified by SIC codes, with the addition of a "software" division, which we carve out of the services division and define through SIC codes 7370 - 7377. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we define “size decile” as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the firm level.

Finally, we consider using a larger and more inclusive definition of "firms". Per Appendix C, we define firms in our baseline sample by requiring that there be either internal or external information on the firm's industry and country. In this larger sample, we drop the restriction that we must observe the firm's industry. Because there is no industry information, we amend the specification in equation (2) so that fixed effects do not vary by industry. Table OA-9 below presents our estimates using this alternative sample.

Table OA-9: Short- and Long-Run Effects of the GDPR
(More Inclusive Definition of Firms)

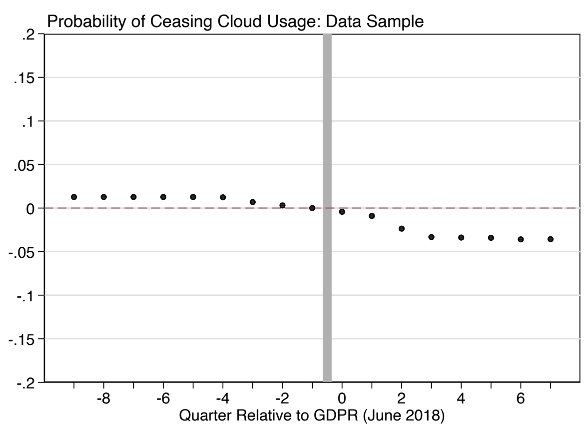
	Data (1)	Compute (2)	Data Intensity (3)
Short-Run Effect	-0.073	-0.059	-0.063
	(0.013)	(0.013)	(0.015)
Long-Run Effect	-0.151	-0.113	-0.117
	(0.018)	(0.020)	(0.022)
Observations	2,224,810	1,097,922	756,996
US Firms	34,876	18,037	10,807
EU Firms	31,622	15,004	10,299

Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. However, we do not allow the fixed effects to vary across industries (not all firms have industry information). Column (1) estimates the effect on storage. Column (2) estimates the effect on computation. Column (3) presents estimates of the data intensity. The sample incorporates firms for which we do not observe industry information, as described in Appendix D. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we define "size decile" as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the firm level.

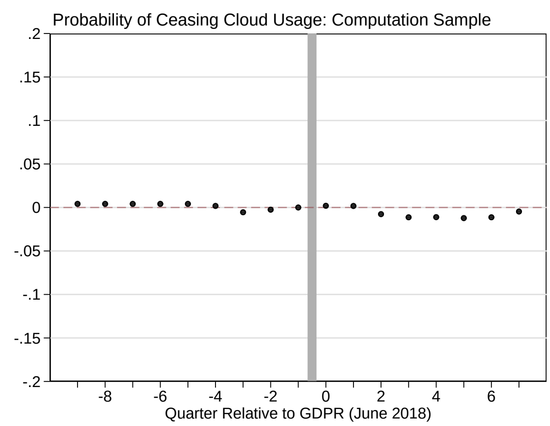
Extensive Margin Although Table OA-3 suggests that our baseline estimates are similar when we use a balanced panel of firms, we also directly examine whether the GDPR caused differential attrition between firms in the European Union and the United States. We study this using the following same specification but replacing the outcome variable with an indicator for whether the firm has exited our sample. We present these results in Figure OA-13.

Figure OA-13: Event Study Estimates of the Effects of the GDPR on Cloud Inputs (Differential Attrition)

(a) Data Sample



(b) Compute Sample



Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. The outcome in each subpanel is denoted by the subpanel title. Dashed bars represent the 95 percent confidence intervals, and standard errors are clustered at the firm level. In contrast to the main figures, the dependent variable is an indicator of whether the firm has exited our sample.

E Technical Appendix

This section provides the derivation of the results in Section 5.

E.1 First-Order Conditions of Cost Minimization

Assume that firms produce according to the following production function:

$$F(X_{it}, I_{it}(C_{it}, D_{it}), \omega_{it}),$$

where I_{it} represents information, X_{it} is a vector of other observed inputs, and ω_{it} represents unobserved productivity. We assume that the information is produced according to the following technology:

$$I_{it} = \left(\omega^c_{it}(C_{it})^\rho + \alpha D^\rho_{it}\right)^{1/\rho}.$$

We assume that firms choose variable inputs to minimize the cost of production by taking prices as given, which is a necessary condition for profit maximization. We also assume that firms take productivity ω^c_{it} as given in the static cost minimization problem. This cost minimization problem can be written as:

$$\min_{C_{it}, D_{it}, X^v_{it}} \quad p^c_{it} C_{it} + p^d_{it} D_{it} + p^x_{it} X^v_{it} \qquad \text{s.t.} \quad F(X_{it}, I_{it}, \omega_{it}) \geq \bar{Y}_{it},$$

where $\bar{Y}_{it}$ is the target level of production, X^v_{it} denotes variable inputs in X_{it}, and p^x_{it} denotes the input price vector of X^v_{it}. The FOCs with respect to C_{it} and D_{it} can be written as:

$$\begin{aligned} \mu_{it} F_2(X_{it}, I_{it}, \omega_{it})\left(\omega^c_{it}(C_{it})^\rho + \alpha D^\rho_{it}\right)^{1/\rho - 1} C^{\rho-1}_{it} w^c_{it} &= p^c_{it} \\ \mu_{it} F_2(X_{it}, I_{it}, \omega_{it})\left(\omega^c_{it}(C_{it})^\rho + \alpha D^\rho_{it}\right)^{1/\rho - 1} D^{\rho-1}_{it} \alpha &= p^d_{it} \end{aligned}$$

where μit is the Lagrange multiplier and F_2 denotes the derivative of F with respect to its second argument. Taking the ratio of the two FOCs, we obtain:

$$\frac{\alpha}{\omega^c_{it}} \left(\frac{C_{it}}{D_{it}}\right)^{1-\rho} = \frac{p^d_{it}}{p^c_{it}}$$

Taking the logarithm and rearranging the terms yields:

$$(1-\rho)\log\left(\frac{C_{it}}{D_{it}}\right) - \log(\omega_{it}^c) + \log(\alpha) = \log\left(\frac{p_{it}^d}{p_{it}^c}\right).$$

By using $\sigma = 1/(1-\rho)$ and defining $\gamma \equiv -\sigma\log(\alpha)$, we can obtain Equation (4) as presented in the main text:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \sigma\log(\omega_{it}^c).$$

E.2 Including Labor in Information Production Function

In this section, we demonstrate that the derivation of the FOCs remains valid even if the information production function includes labor input in the CES form. We consider labor in the information production function because firms might require software engineers to process data. To illustrate this scenario, we consider a nested CES form where data and computation are nested:

$$I_{it} = \left(\left(\omega_{it}^c(C_{it})^\rho + \alpha_D D_{it}^\rho\right)^{v/\rho} + \alpha_L L_{it}^v\right)^{1/v}$$

where v is the substitution parameter between information and labor. Taking the FOCs with respect to C_{it} and D_{it}, we obtain:

$$\mu_{it}F_2(X_{it}, I_{it}, \omega_{it})\left(\left(\omega_{it}^c(C_{it})^\rho + \alpha_D D_{it}^\rho\right)^{v/\rho} + \alpha_L L_{it}^v\right)^{1/v-1}\left(\omega_{it}^c(C_{it})^\rho + \alpha_D D_{it}^\rho\right)^{v/\rho-1} C_{it}^{\rho-1} w_{it}^c = p_{it}^c$$
$$\mu_{it}F_2(X_{it}, I_{it}, \omega_{it})\left(\left(\omega_{it}^c(C_{it})^\rho + \alpha_D D_{it}^\rho\right)^{v/\rho} + \alpha_L L_{it}^v\right)^{1/v-1}\left(\omega_{it}^c(C_{it})^\rho + \alpha_D D_{it}^\rho\right)^{v/\rho-1} D_{it}^{\rho-1} \alpha_D = p_{it}^d$$

Taking the ratio of these FOCs yields the same equation as above:

$$\frac{\alpha_D}{\omega_{it}^c}\left(\frac{C_{it}}{D_{it}}\right)^{1-\rho} = \frac{p_{it}^d}{p_{it}^c}$$

Therefore, the information production function can accommodate labor. It is important to note that this result relies on the specific nested CES functional form used in this analysis. For instance, if data and labor were in the same nest with computation in a different one, the ratio of FOCs would involve labor, and our equivalence result would break down.

E.3 Extensions to the GDPR as a Cost Shock to Data

In this section, we show how our estimates of the wedge induced by the GDPR (λ) would change under alternative assumptions about how the GDPR impacts firms' information production functions. This section builds on details of our identification and estimation procedure described in Section 5.3.

E.3.1 Existing Pre-GDPR Wedges

First, we consider the case in which there are other unobserved variable costs to using data that generate wedges even before the GDPR. In this case, our estimates capture the *additional* wedges driven by the GDPR between the marginal product of data and its price. In particular, consider the following model of data costs faced by each firm i:

$$\textbf{Pre-GDPR:} \quad \tilde{p}^d_{it} = (1+\lambda^0_i)p^d_{it}, \qquad\qquad \textbf{Post-GDPR:} \quad \tilde{p}^d_{it} = (1+\lambda^1_i)p^d_{it}.$$

Under this assumption, our pre-GDPR equation–from which we estimate the firm-specific compute augmenting productivity–becomes:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma_1 \log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma_1\Big(\log(\omega^c_i) + \log(1+\lambda^0_i)\Big) + \sigma_1 \log(\phi^c_t) + \sigma_1 \log(\eta_{it}),$$

so our first-step estimation cannot separately identify ω^c_i from λ^0_i (our estimating equation recovers $\log(\omega^c_i) + \log(1+\lambda^0_i)$ instead of $\log(\omega^c_i)$ under the paper main assumptions).

In the post-GDPR period, the FOCs will be given by:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma_2 + \sigma_2\Big(\log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \log(\phi^c_t)\Big) + \sigma_2\Big(\log(1+\lambda^1_i) + \log(\omega^c_i)\Big) + \sigma_2 \log(\eta_{it}),$$

So in the second step, our estimation procedure recovers $\log(1+\lambda^1_i) + \log(\omega^c_i)$ as the fixed effects. In order to identify the wedges, we subtracted the first-step firm-fixed effect estimates (which estimates $\log(\omega^c_i)$ under our assumptions) from the second-step fixed effects. However, in the model described in this section, the first step recovers an estimate of $\log(\omega^c_i)+\log(1+\lambda^0_i)$ as firm fixed effects. Therefore, subtracting the second-step estimate $\log(\omega^c_i) + \log(1+\lambda^0_i)$ from the second step estimate $\log(1+\lambda^1_i) + \log(\omega^c_i)$ will yield:

$$\log(1+\lambda^1) - \log(1+\lambda^0)$$

Therefore, our procedure recovers $(1+\lambda^1)/(1+\lambda^0) - 1$ as the wedge under the model described in this section, which is the additional multiplicative wedge due to the GDPR.

E.3.2 Negative Productivity Shock to Data-Augmenting Productivity

Our main text assumes that the production function has compute-augmenting productivity. Here, we consider an alternative assumption that productivity is data-augmenting:

$$I_{it}(C_{it}, D_{it}) = \left(\alpha(C_{it})^{\rho} + \tilde{\omega}^{d}_{it} D^{\rho}_{it}\right)^{1/\rho},$$

where $\tilde{\omega}^{d}_{it}$ denotes data-augmenting productivity, which can potentially be affected by the GDPR. In particular, we specify $\tilde{\omega}^{d}_{it}$ as:

$$\textbf{Pre-GDPR:} \quad \tilde{\omega}^{d}_{it} = \omega^{d}_{it}, \qquad\qquad \textbf{Post-GDPR:} \quad \tilde{\omega}^{d}_{it} = (1+\lambda^{d}_{i})\omega^{d}_{it}.$$

where ω^{d}_{it} is the counterfactual data-augmenting productivity in the absence of the GDPR. Here, $\lambda^{d} \leq 0$ corresponds to a negative productivity shock to data-augmenting productivity. Under these assumptions, the FOC in the pre-GDPR period becomes:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma \log\left(\frac{p^{d}_{it}}{p^{c}_{it}}\right) - \sigma \log(\omega^{d}_{it}).$$

Therefore, our first stage procedure recovers firm-specific data-augmenting productivity (formally $-\log \omega^{d}_{it}$) instead of compute-augmenting productivity ($\log \omega^{c}_{it}$). Under the assumption that the GDPR affects the productivity of data, the production function in the post-GDPR period becomes

$$I_{it}(C_{it}, D_{it}) = \left(\alpha(C_{it})^{\rho} + \omega^{d}_{it}(1+\lambda^{d}_{i}) D^{\rho}_{it}\right)^{1/\rho}, \tag{14}$$

Taking the FOCs after the GDPR, we obtain

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma \log\left(\frac{p^{d}_{it}}{p^{c}_{it}}\right) - \sigma \log(\omega^{d}_{it}) - \sigma \log(1+\lambda^{d}_{i}). \tag{15}$$

Compare Equation (15) with our post-GDPR FOC in the main text (Equation (9) reproduced below without the changes in the elasticity of substitution for simplicity):

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma \log\left(\frac{p^{d}_{it}}{p^{c}_{it}}\right) + \sigma \log(1+\lambda_{i}) + \sigma \log(\omega^{c}_{it}).$$

Since we recovered $-\sigma \log(\omega^{d}_{it})$ in the first step, our estimation procedure recovers $\sigma \log(1+\lambda_i)$ as $-\sigma \log(1+\lambda^{d}_{i})$ from Equation (15). Now we can solve for λ_i as a function of λ_i^d from the relationship $1 + \lambda_i = 1/(1 + \lambda_i^d)$ which yields:

$$\lambda_i = \frac{-\lambda_i^d}{1 + \lambda_i^d},$$

and for small λ_i^d, we obtain $\lambda_i \approx -\lambda_i^d$ so our procedure recovers the magnitude of the shock to data productivity due to GDPR. For larger values of λ_i, we can use the exact formula to estimate changes. For example, in the paper, we estimate $\lambda_i \approx 1/5$ on average, which implies that $\lambda_i^d \approx -1/6$ under this alternative assumption.

E.3.3 Wedges in Both Data and Computation

In our main text, we assume that GDPR only affects data costs. Here, we consider the case in which the GDPR affects both computation and data so that:

$$\textbf{Pre-GDPR:} \quad \tilde{p}_{it}^d = p_{it}^d, \quad \tilde{p}_{it}^c = p_{it}^c, \quad \textbf{Post-GDPR:} \quad \tilde{p}_{it}^d = (1 + \lambda_i^d) p_{it}^d, \quad \tilde{p}_{it}^c = (1 + \lambda_i^c) p_{it}^c.$$

Taking first-order conditions post-GDPR under this assumption, we obtain:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma_2 \log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \sigma_2 \log\left(\frac{1 + \lambda_i^d}{1 + \lambda_i^c}\right) + \sigma_2 \log(\omega_i^c) + \sigma_2 \log(\phi_t^c) + \sigma_2 \log(\eta_{it}),$$

which differs from the main text (Equation 9) only through having the ratio of wedges instead of data wedge. Therefore, in this case, our estimation procedure identifies $(1 + \lambda_i^d)/(1 + \lambda_i^c)$ instead of only $(1 + \lambda_i^d)$, which introduces downward bias for λ_i^d if $\lambda_i^c > 0$. Since λ_i^d would be underestimated while we assume $\lambda_i^c = 0$, we would underestimate the total cost of GDPR.

E.4 Derivation for Cost of Information

In this subsection, we derive the formula for the cost of information given by Equation (11) in the paper. Next, we generalize the cost of information for any monotonic transformation of the information production function (e.g., assuming increasing/decreasing returns to scale in the information production function). We then conclude by showing how the percentage changes in the cost of information, as computed in the paper, are invariant to monotonic transformations. To ease notation, we drop the subscripts and use p_c, p_d and ω in the place of p^c, p^d and ω^c.

Consider the optimal isocline as given by the FOC from data and computation:

$$\frac{\alpha}{\omega}\left(\frac{C}{D}\right)^{1-\rho} = \frac{p_d}{p_c} \qquad \Longleftrightarrow \qquad D = \left(\frac{\omega}{p_c}\frac{p_d}{\alpha}\right)^{\frac{1}{\rho-1}} C, \tag{16}$$

which relates the optimal data input usage D to the optimal computation usage C.

To obtain the information production as a function of parameters, we substitute Equation (16) into the information production function:

$$I = \left(\omega C^{\rho} + \alpha\left(\frac{\omega}{p_c}\frac{p_d}{\alpha}\right)^{\frac{\rho}{\rho-1}} C^{\rho}\right)^{1/\rho} = \left(\frac{\omega}{p_c}\right)^{\frac{1}{\rho-1}}\left(p_c\left(\frac{p_c}{\omega}\right)^{\frac{1}{\rho-1}} + p_d\left(\frac{p_d}{\alpha}\right)^{\frac{1}{\rho-1}}\right)^{1/\rho} C \tag{17}$$

and defining Φ as:

$$\Phi = \left(p_c\left(\frac{p_c}{\omega}\right)^{\frac{1}{\rho-1}} + p_d\left(\frac{p_d}{\alpha}\right)^{\frac{1}{\rho-1}}\right),$$

we can simplify Equation (17) as:

$$I = \Phi^{1/\rho}\left(\frac{\omega}{p_c}\right)^{\frac{1}{\rho-1}} C \implies C^{*}(I,p) = \frac{I}{\Phi^{1/\rho}}\left(\frac{p_c}{\omega}\right)^{\frac{1}{\rho-1}} \text{ and } D^{*}(I,p) = \frac{I}{\Phi^{1/\rho}}\left(\frac{p_d}{\alpha}\right)^{\frac{1}{\rho-1}}$$

so we obtain the optimal input demands as a function of prices and parameters. Now, substituting them into the cost of information:

$$CI^{*}(I,p) = p_c C^{*} + p_d D^{*} = \frac{I}{\Phi^{1/\rho}}\underbrace{\left(p_c\left(\frac{p_c}{\omega}\right)^{\frac{1}{\rho-1}} + p_d\left(\frac{p_d}{\alpha}\right)^{\frac{1}{\rho-1}}\right)}_{\phi} = I\Phi^{\frac{\rho-1}{\rho}}.$$

To get to the final result, note that $(\rho-1)/\rho = 1/(1-\sigma)$, and $1/(\rho-1) = -\sigma$. Therefore, we can express the cost of information as a function of I, prices, and parameters:

$$CI^{*}(I,p) = I\left(\omega^{\sigma} p_c^{1-\sigma} + \alpha^{\sigma} p_d^{1-\sigma}\right)^{1/(1-\sigma)},$$

which is the main equation in the paper.

Next, we derive the cost of information for any monotonic transformation of the production function of I. As we argued in Section 5.1, I does not have a natural scale and can be defined only up to a monotonic transformation. For any monotonic transformation of

I, $h(I)$, the cost function can be obtained as:

$$CI^*(I_{it}, p_{it}) = h^{-1}(I)\big(\omega^\sigma p_c^{1-\sigma} + \alpha^\sigma p_d^{1-\sigma}\big)^{1/(1-\sigma)}. \tag{18}$$

To see this, note that when substituting Equation (16) into the production function we get Equation (17)′ as given by:

$$h^{-1}(I) = \left(\frac{\omega}{p_c}\right)^{\frac{1}{\rho-1}} \left(p_c \left(\frac{p_c}{\omega}\right)^{\frac{1}{\rho-1}} + p_d \left(\frac{p_d}{\alpha}\right)^{\frac{1}{\rho-1}}\right)^{1/\rho} C,$$

while the rest of the algebra stays the same but replacing I for $h^{-1}(I)$.

Finally, to show that the *percentage* change in the cost of information is invariant to monotonic transformations, notice that at any information level I, we can take the ratio of Equation (18) with and without the GDPR wedge and subtract one to obtain the percentage change in the cost of information:

$$1 + \Delta CI^*(I_{it}, p_{it}) = \left[\frac{\omega^\sigma p_c^{1-\sigma} + \alpha^\sigma(1+\lambda_i)p_d^{1-\sigma}}{\omega^\sigma p_c^{1-\sigma} + \alpha^\sigma p_d^{1-\sigma}}\right]^{1/(1-\sigma)}, \tag{19}$$

which is the main formula we use in the paper.

E.5 Cost of Information Decomposition

In this section, we derive the formula for the decomposition of the cost of information. We drop all subscripts to ease notation and start by substituting the values for the cost-minimizing information cost, CI^*, as:

$$CI^*(I, p, \lambda) = p_c C^*(I, p, \lambda) + p_d D^*(I, p, \lambda),$$

where $C^*(I, p, \lambda)$ and $D^*(I, p, \lambda)$ are the optimal compute and data choices as a function of information level, input prices, and wedges. We will remove the function arguments to ease out notation even more. The total derivative with respect to λ is obtaining by differentiating both sides with respect to λ:

$$\frac{\mathrm{d}CI^*}{\mathrm{d}\lambda} = p_c\frac{\mathrm{d}C^*}{\mathrm{d}\lambda} + p_d D^* + p_d(1+\lambda)\frac{\mathrm{d}D^*}{\mathrm{d}\lambda}.$$

Multiplying both sides by λ/CI^* we obtain:

$$\frac{\mathrm{d}CI^*}{\mathrm{d}\lambda}\frac{\lambda}{CI^*} = p_c \frac{\mathrm{dC}^*}{\mathrm{d}\lambda}\frac{\lambda}{C^*} + \lambda\left(\frac{p_d D^*}{CI^*}\right) + p_d(1+\lambda)\frac{\mathrm{dD}^*}{\mathrm{d}\lambda}\frac{\lambda}{C^*}.$$

Rearranging terms and multiplying the first term by C^*/C^*, and the third by D^*/D^* we get

$$\frac{\mathrm{d}CI^*}{\mathrm{d}\lambda}\frac{\lambda}{CI^*} = \lambda\left(\frac{p_d D^*}{CI^*}\right) + \left(\frac{p_c C^*}{CI^*}\right)\left[\frac{\mathrm{dC}^*}{\mathrm{d}\lambda}\frac{\lambda}{C^*}\right] + \left(\frac{p_d(1+\lambda)D^*}{CI^*}\right)\left[\frac{\mathrm{dD}^*}{\mathrm{d}\lambda}\frac{\lambda}{D^*}\right],$$

and finally, recognizing that the terms in parenthesis are the expenditure shares s_d and s_c, and the terms in squared parenthesis are the elasticities, we get to the following equation:

$$\varepsilon(CI^*_{it}, \lambda_i) = s^d_{it} \cdot \lambda_i + \left[s^d_{it} \cdot \varepsilon(D^*_{it}, \lambda_i) + s^c_{it} \cdot \varepsilon(C^*_{it}, \lambda_i)\right].$$

F Model Estimation Details

This section provides details on cloud computing pricing, the instrumental variable strategy, our estimation procedure, and intuition for our identification.

F.1 Cloud Computing Pricing

Our estimation of the elasticity of substitution is identified by how firms adjust their input demand to price changes. To provide context for the main sources of price variation, this section presents an overview of pricing in cloud computing.

Cloud computing providers typically consider various factors when choosing cloud prices in different locations. Some of these factors may include the cost of electricity, the availability of skilled labor, the cost of real estate, tax incentives, regulatory requirements, and the availability and cost of network connectivity. Additionally, firms may consider the level of competition in each location and the pricing strategies of different cloud providers.

The pricing of cloud services in the last decade has been characterized by a steady decline across all providers. As cloud providers have achieved economies of scale and improved their technological infrastructure, they have been able to offer lower prices to customers. In addition, increased competition among cloud providers in attracting customers has also contributed to lower prices. Byrne et al. (2018) constructs a price index for AWS over the last decade and investigates how prices have evolved. They found that AWS computation prices fell at an average annual rate of about 7%, database prices fell at an average annual rate of more than 11%, and storage disk prices fell at an annual rate of more than 17%. Part of this price decline is driven by competition. Byrne et al. (2018) finds that AWS prices dropped more significantly when Microsoft Azure entered the market, at 10.5%, 22%, and about 25% for computation, database, and storage, respectively, between 2014 and 2016.

Overall, the last decade has seen a notable trend of declining cloud prices despite increasing demand. This suggests that factors such as competition and technological advances have been the major drivers of cloud pricing in the last decade.

F.2 Price Index Construction

Our instrumental variable strategy relies on constructing firm- and location-specific price indices. This section describes how we construct those price indices.

To obtain firm-specific price indices, we simply calculate the unit price paid by the firm by dividing the monthly total spending on compute and storage by the total quantity of compute and storage, respectively. This gives us firm-specific compute and storage price indices, which can vary either because of discounts due to long-term commitments or variations in location-specific prices. We divide the price of storage by the price of computation to obtain a firm-specific storage-to-compute price ratio. Since this ratio involves some outliers due to small values in the dominator, we winsorize these variables by the top and bottom two percentiles. We also construct the storage-to-compute usage ratio for each firm and apply the same winsorization procedure.

We also calculate location-specific price indices for computation and storage for our sample period. An important issue to account for when calculating these price indices is the entry and exit of products. All cloud providers have introduced a variety of products in the last decade. We construct the price index in the following manner: for any given data location, we first identify products that are available in two adjacent periods, t and $t+1$. We then use the following formula to calculate the price change in location l:

$$r^j_{lt} = \frac{\sum_i p^j_{il(t+1)} q^j_{ilt}}{\sum_i p^j_{ilt} q^j_{ilt}},$$

where $j \in \{c, d\}$ denoting computation and storage, q^j_{ilt} is the total quantity of product i in location l at time t. We calculate this price change for every location-month combination in our sample and construct a price index by cumulatively multiplying the changes in the price index, that is $p^j_{lt} = \prod_{1 \le j \le t} r^j_{lj}$, where $j \in \{c, d\}$ denoting computation and storage.

F.3 Instrumental Variable Strategy

Our instrumental variable strategy relies on the assumption that firms' choice of data center location is persistent. This assumption is based on the fact that the cost of moving large datasets from one data center to another is typically high. The cost of moving data to another data center in cloud computing can depend on several factors, including the amount of data being transferred, the distance between the source and destination data centers, and the pricing policies of the cloud service provider (García-Dorado and Rao, 2015). Some cloud service providers may charge a fee for data transfer, and there may be additional costs associated with data migration, such as network bandwidth charges, storage costs, and downtime or disruption to services during the migration process.[51] Even though the specific costs and risks of data migration will depend on the migration plan and the cloud service provider, it is typically considered too costly by industry experts.

[51] See detailed information on data transfer costs for top cloud computing providers at AWS Data Transfer Costs, Azure Bandwidth Pricing, and Google Cloud Storage Transfer Pricing.

We use the persistence in data center location that comes from switching costs to design a shift-share instrumental variable strategy. Formally, each firm has exposure to different locations and pays different prices in each location due to variations in list prices and firm-specific discounts. We denote firm-specific price indices by p^d_{it} and p^c_{it} for data and computation, respectively. This price could be endogenous because the firm may receive discounts due to long-term commitments or change its exposure to different locations based on productivity. To instrument for these prices, we use the list prices of storage in location l, given by p_{lt}. This price is plausibly exogenous to changes in firm productivity because, after controlling for industry-specific trends, no firm is likely to affect list prices in a specific location. Additionally, we attempt to purge these shares of endogeneity further by taking lags, as contemporary shares may be susceptible to reverse causality. Hence, our instrument for data is given by $z^d_{it} = \sum s^d_{il(t-12)} p^d_{lt}$ for storage and z^c_{it} for computation calculated similarly. Finally, we use z^c_{it}/z^d_{it} to instrument for p^c_{it}/p^d_{it} in the production function estimation. Since we need the 12 months of lagged exposure of each firm, we lose the first 12 months of observations when implementing this instrumental variable strategy.

F.4 Estimation Details

Our identification strategy relies on the assumption that the industry-specific cloud productivity trend in the EU would have followed that of US firms in the absence of the GDPR and that firm-specific compute technology does not change post-GDPR. To operationalize these assumptions, we follow a two-step estimation strategy.

In the first step, we estimate the following equation for US firms using the entire sample period with our IV strategy:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma_1^{US} \log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma_1^{US} \log(\omega^c_i) + \sigma_1^{US} \log(\phi^c_t) + \sigma_1^{US} \log(\eta_{it}), \tag{20}$$

When estimating this equation, we normalize γ to zero because it is not separately identified from the mean of ω^c_i. We also normalize ϕ^c_1 to 1 so that the productivity trend is relative to the initial period. Since, by assumption, the US firms have not been exposed to the GDPR, this equation identifies the industry-specific compute-augmenting productivity trends, or $\hat{\phi}^c_t$ in Equation (10). By Assumption (2), the EU industries follow the same trend and we use the estimated $\hat{\phi}^c_t$ for EU firms.[52] Next, we estimate the same equation using EU firms only with pre-GDPR data. This estimation identifies $\hat{\omega}^c_i$ in Equation (10) because there is no distortion before the GDPR to estimate σ_1^{EU}. We report the associated

[52] We also estimate Equation (20) using pre- and post-GDPR data for US firms to separately identify the elasticity of substitution before and after the implementation of GDPR.

elasticity estimates in Figure 4 as the pre-GDPR elasticity of substitution estimates.

These first-step estimations identify provide us with $\hat{\omega}_i^c$ and $\hat{\phi}_t$. Using those we finally estimate Equation (10):

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma_2 + \sigma_2^{EU}\Big(\log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \log(\hat{\phi}_t)\Big) + \sigma_2^{EU}\Big(\log(1+\lambda_i) + \log(\hat{\omega}_i^c)\Big) + \log(\eta_{it}).$$

by constructing the right-hand side variable. We report σ_2^{EU} as the post-GDPR elasticity of substitution estimates in Figure 4. To estimate the wedge, λ_i, we subtract $\log(\hat{\omega}_i^c)$ from the estimated fixed effects in Equation (10) (after accounting for σ_2^{EU}). We report the estimates of λ_i in Figure 5. To account for uncertainty in first-step estimates in standard errors, we follow a bootstrap procedure with 100 repetitions. We resample firms with replacements in each industry-continent group and apply the entire estimation procedure.

We use Equation (11) to estimate the changes in the cost of information, with results reported in Section 6.3. For each estimated ω_i^c, we calculate the cost of information by setting λ_i to its estimated value and 0, which gives us the change in the cost of information due to the GDPR. Since prices and compute-augmenting productivity may change over time, we calculate this change in information cost at every period and report the distribution at the month-firm level in Figure 7(b).

To do the decomposition presented in Equation 6.3, we calculate the cost share of data each period using firms' data input demands and prices. The direct effect is obtained by multiplying data shares with firm-specific wedges. The second term (firm re-adjustment) is obtained by subtracting the direct effect from the change in the cost of information. Similar to the above, we calculate this change in information cost for every period and report the distribution at the month-firm level.

F.5 Identification Intuition for the Firm-Specific Wedges

Having outlined our estimation strategy in the previous subsection, we now explain how our assumptions help us identify the per-firm wedge in the cost of storing data, λ_i. The main goal is to provide intuition on the variation λ_i is intended to capture. We provide intuition for the case where the elasticity of substitution is the same in the EU and in the US (but may vary pre and post-GDPR), as the more general case provides no additional intuition but involves more cumbersome notation. We consciously abuse notation in this section as its main goal is to provide simple equations.

Consider two firms in the same industry, one in the EU (k) and one in the US (j), with the same levels of firm-level compute-augmenting productivity $\omega_k^c = \omega_j^c$. For simplicity (to not carry terms around), assume both firms have the same time-varying shocks (i.e., $\log \eta_{kt} = \log \eta_{jt}$ for all t).[53] Subtracting the pre-GDPR first-order condition (Equation 7) of the US firm from the EU firm equation in a period $\underline{t}$ before GDPR implies that:

$$\Delta_i\left(\frac{C_{i\underline{t}}}{D_{i\underline{t}}}\right) = \sigma_1 \Delta_i\left(\frac{p^d_{i\underline{t}}}{p^c_{i\underline{t}}}\right) \tag{21}$$

where we define $\Delta_i(X_{it})$ as the across-firm (EU vs. US) difference in the logarithm of X_{it} at time t (i.e., $\Delta_i(X_{it}) \equiv \log X_{kt} - \log X_{jt}$). Note that Assumption 2 (i.e., EU and US industries follow the same compute augmenting productivity time trend) allows us to get rid of ϕ^c_t if we look at two firms within the same period t. Similarly, by focusing on comparable firms (k and j), we get rid of ω^c_k and ω^c_j.

Analogously, focusing on a period $\bar{t}$ after GDPR was enacted, we can use the post-GDPR identifying equation (Equation 9) in a similar fashion as before (focusing on the same two firms) to obtain:

$$\Delta_i\left(\frac{C_{i\bar{t}}}{D_{i\bar{t}}}\right) = \sigma_2 \Delta_i\left(\frac{p^d_{i\bar{t}}}{p^c_{i\bar{t}}}\right) + \sigma_2 \log(1+\lambda_i) \tag{22}$$

where the extra term is the increase in the cost (λ_i) incurred by the firm in the EU but not by the firm in the US. Subtracting both equations, rearranging terms, and some algebra, we get:

$$\Delta\Delta_{it}\left(\frac{C_{it}}{D_{it}}\right) = \sigma_2 \Delta\Delta_{it}\left(\frac{p^d_{it}}{p^c_{it}}\right) + (\sigma_2 - \sigma_1)\Delta_i\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma_2 \log(1+\lambda_i) \tag{23}$$

where $\Delta\Delta_{it}(X_{it})$ is the double difference across the EU and US firms and before and after GDPR (i.e., $\Delta\Delta_{it}(X_{it}) \equiv \Delta_i(X_{i\bar{t}}) - \Delta_i(X_{i\underline{t}})$ in our case). These double differences are akin to the ones one would need to generate a difference in difference estimate (e.g., to those in Section 4 of the paper).

Equation (23) provides useful intuition about what λ_i, the post-GDPR wedge, is intended to capture. Loosely speaking, the wedge captures the variation in the shift in the compute intensity (across EU and US firms, before and after GDPR) that is not explained by changes in the shift in the relative prices or by pre-and post-GDPR differences in the elasticity of substitution between compute and storage across comparable EU and US firms.[54] Given the above equation, one would intuitively expect firms that face larger

[53] Otherwise, we can work with expectations and use precise (but somewhat cumbersome) notation.

[54] The more general case that we estimate, where the elasticity of substitution differs between EU and US firms, has a similar intuition but also involves the difference in the changes in σ between the US and the EU, before and after GDPR. We estimate that these differences are not economically important in our context.

changes in the compute intensity (the negative of the data intensity) to be those that have larger wedges.

Reassuringly, the intuition we explain above is also consistent with our estimated wedges. Recall that we show in the paper that firms became less data-intensive (equivalently, more compute-intensive) after the GDPR. Importantly, we show that industries with larger changes in data intensity are those with larger wedges. Panel C of Table 4 shows that the changes in the data intensity are smaller (in absolute value) for manufacturing firms, followed by firms in the services industry, and then by software firms. Similarly, our average wedge estimates (shown in Figure 5) have the same ordering: manufacturing firms face smaller wedges, followed by services, and finally by software.

Interestingly, Equations (23) and (22) also show that level changes in C_{it} and D_{it} are not enough to identify λ_i. Note that we cannot infer that firms with larger responses in *levels* would have larger (or smaller) wedges. In fact, to rationalize the level of responses to computing and storage, one would need additional assumptions about the full production function. To explain the responses in levels, we would need to construct a model that incorporates the elasticity of substitution between information and other traditional inputs (e.g., capital and labor).

G Effects on Production Costs

G.1 The Effect of Changes in Information Costs on Production Costs

In this section, we consider how changes in information costs translate into changes in production costs under various benchmark production function specifications. Per Section 6.4, this exercise aims to derive simple sufficient statistics under various functional form assumptions for the total increase in the cost of producing goods and services arising from the change in the cost of data storage. As such, we leverage the assumption that firms face linear prices (p) for all inputs. Thus, the resulting cost function is given by:

$$C\left(\bar{Y}, p, \Delta CI\right) = p_L L^*\left(\bar{Y}, p, \Delta CI\right) + p_K K^*\left(\bar{Y}, p, \Delta CI\right) + p_I I^*\left(\bar{Y}, p, \Delta CI\right).$$

where we use $\bar{Y}$ throughout the section to denote the quantity of production, and where ΔCI is the percentage increase in the information cost.

We first consider two edge cases—Leontief and linear production functions—where information is a perfect complement and a substitute for other inputs. These provide us with intuitive bounds for how changes in the costs of information might translate into production costs. Next, we consider an intermediate case with Cobb-Douglas production technology and derive a simple equation for how changes in information costs translate into production costs after firms re-optimize between inputs. Finally, we analyze a nested CES with information and non-information inputs.

Leontief Production Function

We first consider the simple case of a Leontief production function, where inputs must be combined in fixed proportions:

$$Y = \min\left(\frac{L}{\alpha}, \frac{K}{\beta}, \frac{I}{\gamma}\right).$$

Cost minimization immediately implies that for any given level of production, the input demand functions are given by:

$$L^* = \alpha\bar{Y}, \qquad K^* = \beta\bar{Y}, \qquad I^* = \gamma\bar{Y}.$$

In this case, the cost function is therefore linear in prices, and a ΔCI percentage increase in the cost of information causes an $\Delta CI \cdot s_{it}^{I}$ percentage increase in the cost of production.

Linear Production Function

The case of a linear production function is straightforward, as firms simply choose the most cost-effective input or mix between them if they are equally cost-effective.

$$Y = \alpha L + \beta K + \gamma I.$$

In the interior case where firms were previously producing with non-zero capital or non-zero labor, cost minimization immediately implies that a ΔCI percentage increase in the cost of information translates into a zero percentage increase in the cost of production.

Cobb-Douglas Production Function

Next, we consider the effects of a ΔCI percentage increase in the cost of information for a Cobb-Douglas production function given by

$$Y = L^{\alpha} K^{\beta} I^{\gamma}$$

First-order conditions imply the following information demand function:

$$I^* = \bar{Y}^{\frac{1}{\gamma+\alpha+\beta}} \cdot \left(\frac{p^I}{\gamma}\right)^{\frac{-\alpha-\beta}{\gamma+\alpha+\beta}} \cdot \left(\frac{\beta}{p^K}\right)^{\frac{-\beta}{\gamma+\alpha+\beta}} \cdot \left(\frac{\alpha}{p^L}\right)^{\frac{-\alpha}{\gamma+\alpha+\beta}}$$

This immediately implies that a ΔCI percentage increase in p^I induces a $\delta = \left[(1+\Delta CI)^{-\frac{\alpha+\beta}{\gamma+\alpha+\beta}} - 1\right]$ percentage decrease in I^*.[55] Next, we note that first-order conditions imply that a γ share of total firm costs will be spent on information:

$$\gamma = \frac{p^I \cdot I^*\left(\bar{Y}, p, \Delta CI\right)}{E\left(\bar{Y}, p, \Delta CI\right)}.$$

Using the change in information expenditure resulting from the ΔCI increase in information prices and the δ decrease in I^* derived above, we have that a ΔCI percentage increase in p^I will lead to a ΔC percentage increase in production costs, where $\Delta C = (1+\Delta CI)^{\gamma} - 1$.[56]

[55] For marginal changes, using log transformations and taking derivatives yields $\frac{\partial \log I}{\partial \log p^I} = \frac{\alpha+\beta}{\gamma+\alpha+\beta}$.

[56] Once again using log transformations and taking derivatives yields the intuitive expression $\frac{\partial \log(E)}{\partial \log(p^I)} = 1 - \frac{\alpha+\beta}{\gamma+\alpha+\beta}$ for marginal changes from $\Delta CI = 0$.

CES Production Function

Finally, we consider a simple nested constant elasticity of substitution production technology, where information I is combined with constant returns to scale aggregator of all non-information inputs $M(L, K)$. We denote the outer nest by

$$Y_i = \nu_i \left(\beta I_i^{\bar{\rho}} + (1-\beta) M_i^{\bar{\rho}})\right)^{1/\bar{\rho}} .$$

where ν_i represents firm-specific productivity, a_i represents firm-specific information intensity in production, and $\bar{\sigma} = 1/(1-\bar{\rho})$ denotes the elasticity of substitution between information and non-information inputs. Moving forward, we will drop the firm-specific subscripts for notational simplicity.

Next, we note that because $M(L, K)$ exhibits constant returns to scale, the linear prices of labor and capital – p_L and p_K – imply a linear unit cost for the intermediate non-information aggregate M. We denote that unit cost by p_M.[57] This, therefore, yields the unit cost function

$$c(p_I, p_M) = \frac{1}{\nu} \left(\beta^{\bar{\sigma}} (p_I)^{1-\bar{\sigma}} + (1-\beta)^{\bar{\sigma}} (p_M)^{1-\bar{\sigma}}\right)^{\frac{1}{1-\bar{\sigma}}} .$$

Now, denote the equilibrium information expenditure share as $s_I^* \equiv \frac{p_I \cdot I}{p_M \cdot M + p_I \cdot I}$. Combining this with first-order conditions allows us to express this term as

$$\frac{s_I^*}{1 - s_I^*} = \left(\frac{p_I}{p_M}\right)^{1-\bar{\sigma}} \left(\frac{\beta}{1-\beta}\right)^{\bar{\sigma}} .$$

Finally, we can use this equivalence to express the effects of a ΔCI percentage increase in p_I on production costs using only model parameters and s_I^*:

[57] Deriving the formula for the unit cost of M yields $p_M = \frac{1}{\gamma} \left(\beta_{kl}^{\sigma} p_L^{(1-\sigma_{kl})} + (1-\beta)_{kl}^{\sigma} p_K^{(1-\sigma_{kl})}\right)^{1/(1-\sigma_{kl}}$ where σ_{kl} denotes the elasticity of substitution between capital and labor.

$$
\begin{aligned}
\frac{c\big((1+\Delta CI)p_I, p_M\big)}{c\,(p_I, p_M)} &= \left(\frac{(1+\Delta CI)^{1-\bar{\sigma}}\beta^{\bar{\sigma}}p_I^{1-\bar{\sigma}} + (1-\beta)^{\bar{\sigma}}p_M^{1-\bar{\sigma}}}{\beta^{\bar{\sigma}}p_I^{1-\bar{\sigma}} + (1-\beta)^{\bar{\sigma}}p_M^{1-\bar{\sigma}}}\right)^{\frac{1}{1-\bar{\sigma}}} \\
&= \left(\frac{(1+\Delta CI)^{1-\bar{\sigma}}\left(\frac{\beta}{1-\beta}\right)^{\bar{\sigma}}\left(\frac{p_I}{p_M}\right)^{1-\bar{\sigma}} + 1}{\left(\frac{\beta}{1-\beta}\right)^{\bar{\sigma}}\left(\frac{p_I}{p_M}\right)^{1-\bar{\sigma}} + 1}\right)^{\frac{1}{1-\bar{\sigma}}} \\
&= \left((1+\Delta CI)^{1-\bar{\sigma}}s_I^* + 1 - s_I^*\right)^{\frac{1}{1-\bar{\sigma}}}.
\end{aligned}
$$

Thus, a ΔCI percentage increase in p_I yields a $\left((1+\Delta CI)^{1-\bar{\sigma}} \cdot s_I^* + 1 - s_I^*\right)^{\frac{1}{1-\bar{\sigma}}} - 1$ percentage increase in production costs.

G.2 Estimating Key Parameters of Production Cost Increases

We show in the section above that the information share of expenditure is crucial to calculating how an increase in the cost of information translates to production costs. In the nested CES production technology we analyze above, the vector with the elasticity of substitution between information and non-information inputs and the information cost share is a sufficient statistic for this effect. We discuss estimates of both parameters below.

First, we combine various data sources to suggest a reasonable range for the information cost share. We provide these estimates in Table OA-10. Next, we discuss each of those data sources separately. Finally, we discuss mapping estimates from Lashkari et al. (2024) of the elasticity of substitution between IT and non-IT inputs into our setting.

Aberdeen

We begin by turning to the Aberdeen data set, which we discuss in Section 3.2 and in Appendix C.3. The Aberdeen data provides estimates of site-level IT spending and revenue, which we collapse to the firm level. Unfortunately, we are unable to directly observe total firm expenditures, so we proxy instead with firm revenue. We construct the average share of IT revenue spent for European and US firms in 2017 and 2018. We further use the four-digit SIC codes from the data to identify and partition firms that belong to our three primary industries of interest: software, services, and manufacturing. We find that, somewhat unsurprisingly, software firms spend the highest share of their revenue on IT, followed by services and then manufacturing.

Table OA-10: Estimates for the Information Share of Expenditure by Industry

	Software (1)	Services (2)	Manufacturing (3)
		Aberdeen Estimates	
Aberdeen (EU 2017)	16.7%	3.7%	3.3%
Aberdeen (EU 2018)	14.9%	2.9%	2.9%
Aberdeen (US 2017)	8.7%	4.9%	3.0%
Aberdeen (US 2018)	8.7%	5.0%	3.2%
		Survey Estimates	
Flexera (2020)	24.7%	6.7%	4.1%
Gartner (2022)	7.1%	5.4%	2.3%
Computer Economics (2019)	–	–	1.4% - 3.2%

Notes: Table presents estimates for the information share of expenditure by industry. All estimates are formed by calculating or observing the average share of firm revenue spent on IT. Column (1) presents these estimates for software firms, which are defined in the Aberdeen data through SIC codes 7370 - 7377. Column (2) presents estimates for firms in services. Column (3) presents estimates for manufacturing firms. Further details on the Aberdeen data and the survey estimates are provided in Appendix G.2.

Industry Surveys

Next, we use industry surveys as supportive evidence that the ranges suggested by Aberdeen data are reasonable. These surveys include Flexera, Gartner, and Computer Economics. These are specifically Flexera's *2020 State of Technology Spending Report*, Gartner's *IT Key Metrics Data 2023: Industry Measures — Insights for Midsize Enterprises*, and Computer Economics's *2019 IT Spending & Staffing Benchmarks – Executive Summary*. For the Flexera survey, we use the "industrial products" industry estimate as the manufacturing estimate, and for the Gartner survey, we take the "professional services" industry estimate as our estimate for non-software service firms. While the samples and industry definitions vary widely across these surveys, the numbers cited are generally consistent with the ranges suggested by Aberdeen.

Estimates of the Elasticity of Substitution between IT and Non-IT Inputs

We use point estimates of the elasticity of substitution between IT and non-IT inputs from Lashkari et al. (2024) to proxy for the elasticity of substitution between information and non-information inputs. We focus on the micro-elasticities provided in the text rather than the macro-elasticities, which reflect general equilibrium forces and reallocation between firms. We use their industry-level elasticities from a non-homothetic CES specification. Estimates for the manufacturing industry are provided directly. We map the "information and communication technology" industry to software, and we construct an estimate for the elasticity in services by taking a weighted average of the relevant industries for which estimates were provided in the online appendix.

Estimates of the Contribution to GDP by Industry and GDP in the EU Area

To measure each industry's contribution to GDP, we use the information provided by OECD (2020) for the Euro Area using their output approach outlined on page 189. We measure the manufacturing contribution to GDP as the manufacturing output at basic prices (line 4), divided by the total gross value added at basic prices (line 1), and get 16.88%. To measure software and non-services, we use a two-step approach. We first compute the service contribution to GDP by summing all of the service industries in the OECD table (lines 6 to 12) and dividing it by the total gross value added (line 1) to get 73.39%. We then separate into "software" and "non-software" by estimating the share of the software industry as a proportion of the service industry.

To separate these industries, we leverage data from the US census to compute the software industry share of the service sector, as we could not find any reliable estimates for the EU. To compute this, we use the 2019 SUSB Annual Data Tables by Establishment Industry provided by the US Census Office. We compute the software industry share by dividing employment in the software industry by the total employment in the service industry and get 7.53%.[58] We use this number to proxy for the EU size of the software industry.

Finally, we return to OECD (2020) to measure total GDP in the EU area, and we use their estimate for 2018 (line 64 of p. 189), which is €11.5 trillion.

[58]To do this, we map from SIC to NAICS codes using Orbis data and assign each service industry code as "software" or "non-software" to match the definitions used in the paper.