Data, Privacy Laws and Firm Production: Evidence from the GDPR*

Mert Demirer† Diego Jiménez-Hernández‡ Dean Li§ Sida Peng¶

October 30, 2023

Abstract

By regulating how firms collect, store, and use data, privacy laws may change the role of data in production and alter firm demand for computation and data storage. 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 confidential data from one of the world's largest cloud-computing providers. Our difference-in-difference estimates indicate that, in response to the GDPR, EU firms decreased data storage by 26% and processing by 15% relative to comparable US firms, becoming less "data-intensive." To estimate the costs of the GDPR for production, we propose and estimate an "information" production function framework where data and computation serve as inputs to production. We find that data and computation are strong complements in production and that firm responses are consistent with the GDPR representing a 20% increase in the cost of data on average, with smaller firms bearing higher cost increases than larger ones. The production cost of information increased by 4% on average, with higher costs in more data-intensive industries.

JEL: L51, L86, D22, L11

Keywords: privacy laws; production function; GDPR; data; cloud computing

*We thank James Brand, Alessandro Bonatti, Peter Cihon, Joe Doyle, Ben Edelman, Liran Einav, Sara Ellison, Maryam Farboodi, Samuel Goldberg, Garrett Johnson, Gaston Illanes, Markus Mobius, Dominik Rehse, Tobias Salz, Bryan Stuart, Taheya Tarannum, Joel Waldfogel, and Mike Whinston for helpful comments, and Taegan Mullane, Doris Pan, Ryan Perry, Bea Rivera for excellent research assistance. We are also grateful to Han Choi for copyediting assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

†MIT Sloan, mdemirer@mit.edu

‡Federal Reserve Bank of Chicago, diego.jimenezhernandez@chi.frb.org

§MIT Economics, deanli@mit.edu

¶Microsoft, sida.peng@microsoft.com

1 Introduction

In the information age, the economy's 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 on human subjects, its growing use has led to new policy attention and regulation. One of the most influential privacy policies is the European General Data Protection Regulation (GDPR), which was enacted in 2016 and affected more than 20 million firms across dozens of countries (GDPR.eu, 2019). 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, potentially affecting their input choices and production decisions. For example, privacy laws may generate a wedge between the marginal product of data and its (perceived) marginal cost, leading firms to substitute away from data with other inputs. Variations in these wedges across firms can result in input misallocation and aggregate productivity losses (Hsieh and Klenow, 2009; Restuccia and Rogerson, 2017). Given the increasing role of data in firm production, understanding how privacy regulations affect firms' input decisions is therefore of the utmost importance.

Large-scale empirical evidence of how privacy laws affect firm data decisions, the key margin targeted by privacy laws, is scant, as 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. Any such framework needs to be parsimonious while having enough flexibility to allow the impact 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' computation and data choices using confidential data from one of the largest global cloud-computing providers. The cloud is an ideal setting for our question because it allows us to observe high-frequency firm decisions about data and computation usage over a six-year horizon from 2015-2021. Our data contains detailed information on the monthly cloud usage of hundreds of thousands of firms and comprises hundreds of zettabytes (i.e., *hundreds of millions* of terabytes) of data and billions of core-hours.[1] This data spans every top-level industry, from manufacturing to finance, and enables us to analyze the impacts of privacy regulations beyond the digital economy.

[1] We omit precise numbers to avoid disclosing potentially business-sensitive information.

We first apply this data toward studying the direct impact of the GDPR on firm data and computation choices. In our first set of analyses, we compare domestic firms in the European Union (EU) subject to the GDPR to comparable non-treated same-industry firms in the US in a difference-in-differences approach. In the second part of the paper, we develop and estimate a production function framework with data and computation. We use this framework both to study how firms combine data and computation and to infer the wedges generated by the GDPR from the shift in firms' data and computation demand.

We begin by providing an overview of the key features of the GDPR that directly affect firm input decisions. The GPDR is a landmark privacy policy enacted in 2016 and implemented in 2018. Notably, 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 firms to enhance data protection, increasing penalties in case of data breaches, and giving consumers more information about firms' tracking behavior. 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) and one of today's most important methods for data use (Byrne et al., 2018). Using data from our cloud computing provider, we observe firm-level monthly usage of several cloud products, including "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 location of the data centers where firms source services. We match our cloud usage data to other data sources that provide detailed firm characteristics.

Our first set of results comes from an event study design comparing data use and computation among comparable firms in the EU to the US after the GDPR. We find that EU firms store on average 26% less data than US firms two years after the GDPR. The direction of this relative decline in storage is perhaps unsurprising, given that the GDPR primarily regulates data usage, but the magnitude is noteworthy. Interestingly, we also find that EU firms decreased their computation relative to US firms by 15%—a smaller effect than that on data.[2] Thus, firms became less data-intensive after the GDPR, computing with proportionally less data. Furthermore, we observe substantial heterogeneity in the effects of the GDPR across industries. Finally, we look at how the effects vary with the

[2]It is ex-ante unclear how the GDPR would affect computation; this effect theoretically depends on the substitutability between data and computation (Acemoglu, 2002). For example, if data and computation were strong substitutes, firms could replace data with computation to minimize the effects of the GDPR.

regulatory stringency across EU countries as the GDPR is enforced by individual EU countries. Although the differences are not statistically significant, our estimates suggest that firms in countries with stricter regulators respond by decreasing their storage and computation more than those in countries with more lenient ones.

While our reduced form findings provide direct evidence of the impact of privacy laws on firms, they only offer a partial understanding of the associated economic costs. Motivated by this, we propose and estimate a production function model where firms use data and computation to produce "information" through a constant elasticity of substitution (CES) function. This production function includes two main parameters: (i) *the firm-level compute (augmenting) productivity*, which determines relative factor intensities of computation and data (Doraszelski and Jaumandreu, 2018; Raval, 2019; Demirer, 2020) and (ii) *the elasticity of substitution* between computation and data, which determines how firms respond to changes in factor prices (Hicks, 1932). Our model is intentionally agnostic about how information enters the final production function, accommodating several important use cases of data, such as being an intermediate input in the production function and augmenting firm productivity. This model links the theoretical literature of data in the production function (e.g., Jones and Tonetti, 2020; Farboodi and Veldkamp, 2022) with empirical estimates and emphasizes the role of computation in firm production.

Our information production model provides an input demand function that links firms' optimal data and computation choices to input prices and model parameters. We estimate this input demand function industry-by-industry to recover the elasticity of substitution (using pre-GDPR variation) and regulatory wedges (using post-GDPR variation).[3] We estimate that data and computation are strong complements in production, with some heterogeneity across industries. The average elasticity of substitution between storage and computation is 0.41, with estimates ranging from 0.44 (non-software services) to 0.34 (manufacturing). This strong complementarity suggests that firms cannot easily substitute toward computation when faced with increased data costs. To our knowledge, this is the first estimate of the elasticity of substitution between different data inputs.

To recover the distortion generated by the GDPR, we model it as an unobserved wedge between the marginal cost firms must pay to store data in the cloud and the total marginal cost that includes GDPR compliance costs. This wedge arises from various sources, including penalties in case of breaches, higher data security requirements, and the need for detailed data records. We estimate firm-specific wedges by utilizing post-GDPR data and attributing to GDPR-induced wedges the change in input choices unexplained by changes

[3] We also account for potential sources of endogeneity in prices by using a shift-share instrument, which we describe in further detail in Section 5.3.1.

in input prices in the EU (relative to the US), or by changes in the elasticity of substitution.

Our production function analysis suggests that the GDPR made data storage 20% more costly for firms on average. The effect is the largest in the software sector (24%), followed by manufacturing (18%), and services (18%). These results suggest that firms in data-intensive industries face higher costs. What determines the increase in costs? To provide suggestive evidence for this question, we look at what firm-level variables are correlated with the estimated firm-specific wedges. We consider two firm characteristics: (i) firm size, measured by the number of employees, and (ii) pre-GDPR compute productivity, estimated from the production function specification. We find that that larger and more compute-intensive firms experienced smaller wedges from the GDPR.

In the last part of the paper, we use the model to estimate the change in the "cost of information"—the least-costly combination of inputs to produce information given input prices—resulting from the increase in the cost of storing data. We use the model to estimate the difference in the cost with and without the wedge introduced by GDPR holding the price of data and computation and the amount of information produced to post-GDPR levels. We estimate that GDPR made it 4% more costly to produce the same amount of information. Given the strong complementarity of data storage and compute, we decompose the cost of information results and show that firms can only absorb a small share of the increase in costs by reoptimizing data storage and computation inputs. Finally, we conduct a simple back-of-the-envelope calculation assuming Cobb-Douglas production technology in information, labor, and capital to estimate the impact of the increase in the cost of information on production. We estimate increases in production costs on the order of 0.50% for software firms, with smaller effects in the less data-intensive industries.

We conduct additional analyses to show that our results are robust to many concerns, such as observing data from a single cloud provider, endogenous pricing responses by the cloud provider, and changes in website cookie collection behavior. 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 traditional IT is not a large concern. Third, we show that our results are not driven by differential trends in cloud prices in the EU and US. Finally, we estimate our specification while excluding firms using web services or with listed websites, 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 study of a single firm. 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.

We conclude the introduction by highlighting that our results do not provide a definitive answer on the overall welfare impact of privacy laws. On one hand, privacy laws may benefit consumers by protecting their digital privacy (Arrieta-Ibarra et al., 2018). On the other hand, compliance with privacy laws are costly for firms. Our paper presents detailed and large-scale evidence of the costs of privacy laws. However, further evidence is needed to fully understand the benefits of these laws and how they compare with any potential harm to firms.[4]

Contribution to the Literature The first body of literature we contribute to is the research on the impact of the GDPR on firms. These papers find that the GDPR decreased the investment in technology ventures (Jia et al., 2021) while encouraging app exit and discouraging app development (Kircher and Foerderer, 2020; Janßen et al., 2021). Several papers studying the GDPR document adverse impacts on digital tracking and advertising: the GDPR decreased the usage of tracking technology tools, such as cookies, in the immediate months after implementation (Aridor et al., 2022; Lefrere et al., 2022; 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., 2022) and increased search frictions (Zhao et al., 2021). On the benefits side, some papers argue that GDPR requirements may have differentially filtered out low-value customers for firms, increasing the average value of remaining consumers to advertisers (Aridor et al., 2022) and increasing effective targeted advertising (Godinho de Matos and Adjerid, 2022).

A subset of the GDPR papers study outcomes outside the digital economy. These papers find that the GDPR may have decreased profits, sales, and profit margins (Koski and Valmari, 2020; Chen et al., 2022). Some papers were concerned about the effect of privacy regulation on the competitive structure of data-intensive industries, with smaller firms being the most affected (Campbell et al., 2015; Koski and Valmari, 2020). We note that although most evidence suggests that the GDPR has significantly impacted data-driven economic activity, Zhuo et al. (2021) find a null effect for short-term extensive margin changes in the formation and termination of internet infrastructures between GDPR and

[4] The economics of privacy literature consistently finds a discrepancy between individuals' strong stated preferences for privacy and their willingness to share personal information—the "privacy paradox" (Acquisti et al., 2016). This discrepancy makes it particularly challenging to estimate the benefit of privacy.

non-GDPR countries.[5] Johnson (2022) provides a comprehensive survey of this literature.[6]

While our paper builds on an identification strategy similar to some of these GDPR papers, it is different in two main aspects. First, because of the richness of our data, we directly study firms' data and computation decisions, a margin that is the key target of regulation. In particular, our data is well-suited for studying firm adjustments on the intensive margin, and the heterogeneity of our results across industries. Second, we take a production function approach and structurally estimate its parameters. Crucially, this approach allows us to estimate the role of data and computation in production and to calculate the cost of the GDPR for firms.

The second body of literature to which we contribute is the set of papers that include data as an input to production. The theoretical literature on data has proposed ways in which data enters production, mostly including it as an additional input to production. Jones and Tonetti (2020) model data as a non-rival input that is generated as a byproduct of production from all firms in the economy. Farboodi and Veldkamp (2022) model data as a productivity-enhancing input that helps firms accurately predict future outcomes. We complement this literature by developing and estimating a firm production framework with data, providing empirical estimates on how firms combine data and computation.

Third, our paper is related to the literature on misallocation, which documents large differences in the efficiency of factor allocations resulting from various frictions (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). Most of this literature abstracts from the origin of frictions, treating them as model primitives. In contrast, we study an important regulatory change that could impacts firms' input allocation. We employ a similar identification strategy by modeling regulation as a wedge between the marginal revenue product of an input and its price to estimate firm-specific distortions.

Our paper also relates to the growing body of literature on the use of personal data by firms (e.g., Bergemann and Bonatti, 2015; Arrieta-Ibarra et al., 2018; Bergemann et al., 2018; Acemoglu et al., 2022; Bergemann and Bonatti, 2022; Bimpikis et al., 2023) by providing empirical evidence on the value of data for firms. We also directly contribute 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 important firm outcomes.

[5] Our measures, however, enable us to study data-related adjustments by firms more directly and to measure the effects of the GDPR on both intensive margins and over a longer horizon.

[6] More recent literature has studied California Consumer Privacy Act (Canayaz et al., 2022; Doerr et al., 2023).

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 direct impacts on firms but because of its crucial role in shaping privacy laws. In this section, we describe the key features of this policy and how they affect 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.[7] We provide a detailed description of the changes required for firms after GDPR in Appendix B.1 and summarize its most important characteristics below.

There are two aspects of GDPR that are important for our paper and govern our approach to modeling it. First, GDPR takes a data protection approach rather than a consumer protection approach (Jones and Kaminski, 2020).[8] A data protection approach imposes a set of costly responsibilities on firms to protect data, in addition to a substantive system of individual rights. This increases the cost of handling data for firms. Second, GDPR takes a risk-based approach to data protection (Hustinx, 2013; Gellert, 2018). For example, Article 25 (Data Protection by Design and by Default) uses phrases such as "implement appropriate technical and organizational measures," "implement data-protection principles," and "in an effective manner." This risk-based approach makes costs heterogeneous across firms based on the sensitivity of data and firms' risk preferences.

The GDPR applies whenever the firm ("data controller") that controls the data 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). Under the GDPR, personal data is defined broadly to include any information that can be used to identify an individual either directly or indirectly (Article 4). This includes information such as name, address, email address, internet protocol (IP) address, and other identifying

[7]Unlike the GDPR, which is directly binding and applicable across the European Union, the preceding Directive 95/46/EC had to be incorporated into each member state's national laws to take effect, leading to variation in its implementation across different jurisdictions.

[8]Consumer protection approach is the dominant approach in the US (Boyne, 2018).

characteristics. It applies to *all* personal data, regardless of whether it is in a client or employee context. Therefore, even business-to-business firms are subject to GDPR.

From the firm perspective, the GDPR primarily increased the cost of collecting and storing data by imposing costly responsibilities on firms. These include keeping a record of processing activities (Article 30), 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), providing timely notifications in case of data breaches (Article 33), executing consumers' requests for data transfer, erasure, or rectification (Article 14-21), and paying hefty penalties in case of data breaches (Article 83). Firms also must have a legal basis for processing personal data.[9]

The cost of complying with the GDPR can vary significantly 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 for firms. 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).

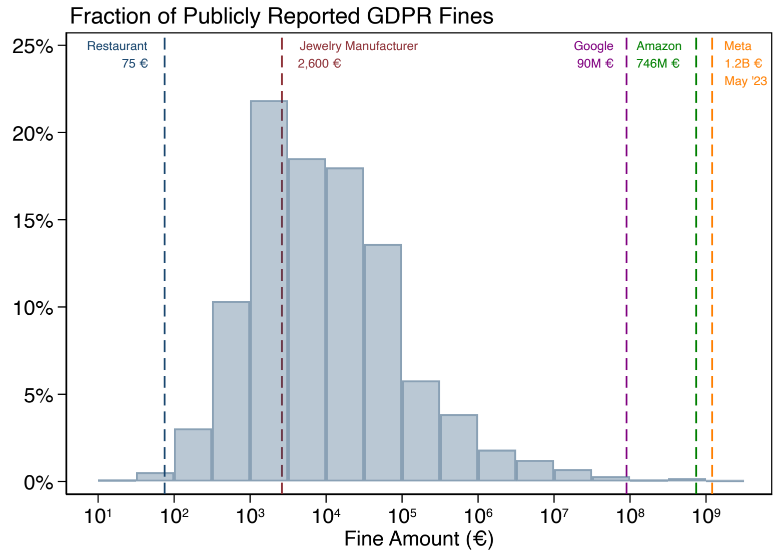
The changes mandated by the GDPR entail both fixed and marginal costs. For example, the cost of having a data protection officer may not scale with data size, so the latter could be considered mostly a fixed cost. On the other hand, the costs of handling customers' access or deletion requests, the liability in case of a data breach, and keeping data in a more secure environment would increase with data and firm size. As such, it may be more sensible to interpret these kinds of costs as changes to the 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, organizations may also incur indirect costs such as cybersecurity insurance or penalties if they are found to be non-compliant or in the case of data leaks.[10] Non-compliant firms may face fines of up to 4% of an organization's annual *global* revenue or €20 million (whichever is greater). We scrape publicly available GDPR

[9]Contrary to popular belief, consent is not the only appropriate legal basis that firms may use to process personal data—consent, contractual necessity, legal obligation, vital interests, public task, and legitimate business interest may all serve as a basis for processing data (Article 6). However, firms are required to identify which legal basis they are using to process personal data.

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

Figure 1: 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. The data collection process is described in Section 3 and we provide greater detail for the data in Appendix B.3. Fines are presented in undeflated nominal terms (€), and five examples from the data have been highlighted: a restaurant, a jewelry manufacturer, Google, Amazon, and Meta.

fine data (which we describe detail in Appendix B.3) from a database maintained by CMS, an international law firm.[11] In Figure 1, we provide the size distribution of these GDPR fines.[12] We note two key features of these fines. First, the distribution of fine sizes implies that enforcement is not limited to large violations: 25% of the fines have been under €2,000. Many of these have been 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 non-software cases, and restaurants and manufacturers appear not infrequently in our dataset.

2.2 Our Setting: Cloud Technology

One of the primary challenges of studying firms' responses to privacy policies has been the fundamental unobservability of how firms use data. Measuring data usage for firms with traditional IT requires both access to their servers and an accounting of usage statistics that firms may not even keep themselves. The advent of cloud computing, however, presents a tremendous opportunity to study the impact of policy changes on firm data usage due to well-tracked measures of storage and processing.

[11] www.enforcementtracker.com.

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

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 is defined as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."[13] Cloud computing has experienced extremely rapid growth since its introduction.[14] According to a 2020 survey by O'Reilly, 88% of respondents used cloud computing in some form.[15]

We focus on the two primary cloud services provided by our data partner: storage and computation. Storage services allow users to store data and applications in a data center location, which can be accessed over the internet. Computation services allow users to run applications and perform computations in a virtual machine (VM). Cloud providers offer a variety of VM types with different specifications in terms of CPU, memory, and upload and download speed. Users choose the VM type that best meets the needs of their workload (Kilcioglu et al., 2017).

Firms could use storage and computing services in multiple parts of their production process. For example, a manufacturing company that produces goods in multiple locations may use VMs to ensure that all of its information is available everywhere (and to monitor inventories, value chains, etc.). Firms may also decide to use storage without using computing services, e.g., a newspaper may decide to host all of the photographs that will be displayed on its website online and provision them directly without the need for computing. However, it is rare to observe firms using computation without also using storage—e.g., some non-data simulations may fit these cases. Firms may also add other cloud services (e.g., analytics, security) in conjunction with their computing and storage needs.[16]

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

[13] Cloud computing resources can be categorized into three forms: Infrastructure as a Service, Platform as a Service, and Software as a Service.

[14] See Jin and McElheran (2017); DeStefano et al. (2020); Jin (2022) for recent studies on firm's cloud adoption and the impacts of cloud technology on firms.

[15] See https://www.oreilly.com/radar/cloud-adoption-in-2020/.

[16] See several case studies of how firms in different industries use cloud computing at https://aws.amazon.com/solutions/case-studies/, https://azure.microsoft.com/en-us/resources/customer-stories/, and https://cloud.google.com/customers.

Despite these advantages, there are important limitations to using data from cloud computing. First, many firms use a mix of cloud computing and traditional 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 limit our analysis if the GDPR changes the composition of cloud and on-site data. Second, it is common for firms to use cloud services from multiple providers, known as multi-cloud. For these firms, a reduction in cloud technology usage from one provider could indicate substitution to another provider. We take these concerns seriously and provide several robustness checks in our empirical strategy.

3 Data

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.

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 usage information of the universe of their customers for all cloud services between 2015 and 2021. These services include hardware services, such as storage, computation, and networking, as well as some software services.[17] For each service, we observe its description, 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.[18]

We focus on storage and computation, as they are the main IT services firms use in cloud computing, which we describe in greater detail in Appendix C.1. We measure storage in gigabytes and computing in core-hours (number of cores $\times$ number of hours). Core-hours are a commonly used metric to quantify the amount of computational work done in cloud computing environments.[19] We use this data to construct monthly-level 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.[20] Through this data, we also observe SIC industry codes, firm headquarters

[17] These software service solutions can be purchased from our provider, but firms may also choose to implement such services themselves manually. In this latter case, we would observe this usage as computation.

[18] This is in contrast with the most input information in production datasets, which generally include input expenditures rather than measures of direct usage.

[19] To illustrate the concept, consider the example of a software engineer in a startup who runs a virtual machine with 8 cores for 5 hours. In this case, the usage is recorded as 40 units of compute.

[20] It is important to note that our sample is comprised of firms rather than establishments.

location, and whether a firm is a start-up or not.[21]

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 Several Providers (2016-2021)

One key concern about using only cloud computing usage data from a single firm is that we cannot observe the margin of usage being diverted to other cloud providers. To address this concern, we use an establishment-level IT data panel produced by a marketing and information company called Aberdeen (previously known as "Harte Hanks"). Using web crawling, surveys and publicly available data, Aberdeen provides the adoption of cloud technology on the extensive margin from each of the service providers (e.g., Amazon, Microsoft, Google) between 2016 and 2021 at the yearly level. The Aberdeen dataset comprises around 3.1 million establishments from 1.9 million companies worldwide. Previous versions of this data have been widely used by researchers to construct measures of IT adoption, both in Europe and in the United States.[22] We use this data to identify single cloud firms and examine differential changes in market share around the GDPR for cloud providers.

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 firms using a matching procedure described in Appendix C.3 based on name, location, domain, and other information. We are able to match close to 60% of our cloud firms to the Aberdeen dataset. We use the employment information in 2018 to define firm size. We further augment our data by merging our primary dataset with Orbis firm database from Bureau van Dijk using firm name and domain name matching. We augment these merges with manual linking for the small share of remaining firms. With this procedure, we link employment data to approximately 80% of the European firms.

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

[22] 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: Matrix of Firms 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.

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 2.1, Table 1 presents information on whether the GDPR applies to firms depending on the location of the firm and their consumers and employees (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 latency is an important feature of cloud usage: the further the distance between the firm and its chosen server, the greater the 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.[23]

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 (Cases 1 and 4) if they use data centers only in Europe or in the US.[24] These domestic firms constitute our main sample throughout the paper.[25]

[23] 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 regions assigned to 98% of the firms coincide with the headquarter locations in our data.

[24] 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.

[25] While multi-national firms are important, their exposure to GDPR and the margins they can respond on vary significantly.

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%

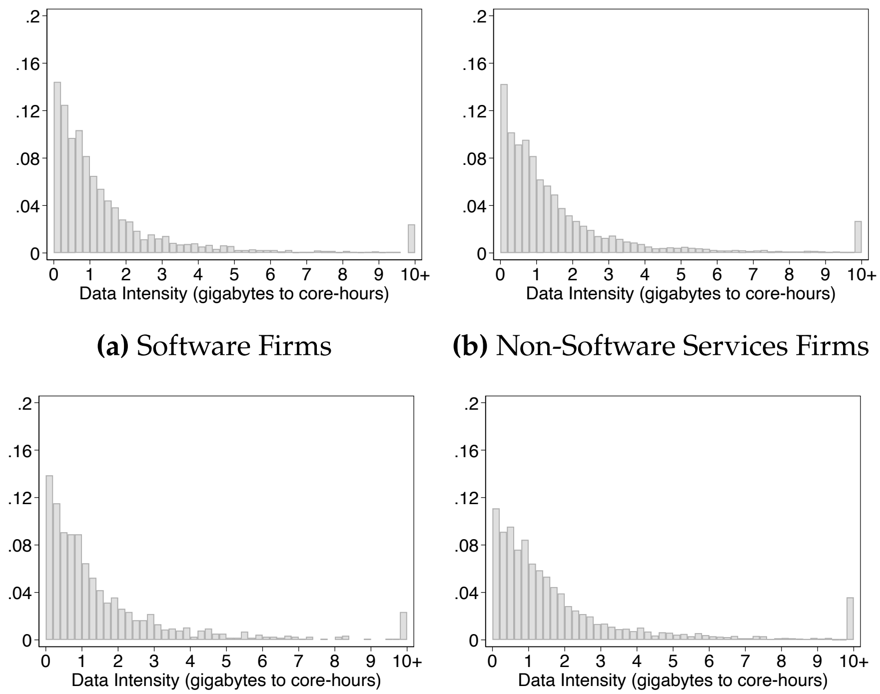
Notes: Table presents summary statistics from our matched sample of firms. A description of the sample's construction can be found in Section 3.1 and a more detailed description of the sample construction can be found in Appendix C. Industries are defined as the ten divisions classified by SIC codes, 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.

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 exactly two years prior to the introduction of the GDPR. This restriction affects only a small share of pre-GDPR storage or computation in our sample: excluded firms are only responsible for about 10% of storage and computation. We use this sample restriction to intentionally focus our analysis on the effects of the GDPR on relatively stable users of our cloud computing provider. 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 the industry of each firm by simply taking the industry division that corresponds to 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.[26] The majority of firms belong to the software (40%) and services (30%) industries, but firms from manufacturing and various other industries are also represented in our sample. 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 that there is some slight variation in the share of firms that we observe in the

[26]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) Non-Software Services Firms

(c) Manufacturing Firms

(d) Other 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, which defined by SIC codes (with the exception of software firms, which are carved out of the services division). We limit to the sample of firms who have ever used both storage and computation ($N = 11,858$).

US versus the EU by industry, although each region always accounts for at least 40% of the share of firms observed.

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 does not vary significantly across industries, ranging from 1.69 to 2.23. However, these averages mask significant within-industry firm-level heterogeneity, as shown in Figure 2, which plots the distribution of data intensity for the three largest industries in our sample. Even within an industry, there is significant firm-level variation in data intensity across all industries, suggesting that the role of data and computation likely vary across firms.[27] 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), markups (Autor et al., 2020; De Loecker et al., 2020), and management practices (Bloom and Van Reenen, 2007). As we will see in Sections 5, taking into account this heterogeneity will be important when modeling a production framework with data and computation.

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

4 Event Study Evidence

In this section, we apply an event study design to study the effect of the GDPR on firms' data storage and computing decisions. We begin by defining our empirical strategy and providing intuition for our identifying assumptions. Next, we turn toward our baseline estimates of the GDPR's impact on data input choices. We also discuss the robustness of our strategy across various alternative samples and specifications. Finally, we estimate how the effects of the GDPR vary across industries in our sample.

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 branches or some of their data may be subject to the GDPR. Thus, we might want to separate the estimation of the treatment effects of these groups of firms from the firms which we consider fully treated (Case 1). Second, multi-national firms may systematically differ from the control firms that we define (Case 4). They may potentially respond to the GDPR, for example, along more margins than our control group, choosing to shift data storage, computation, and even business operations into or out of the European Union.

We focus on three separate outcomes: data storage, computation, and "data intensity" (the ratio of storage to computation). These outcomes reflect the multiple dimensions of firm data usage that might be affected by the GDPR. Our empirical specification uses a difference-in-differences design and estimates the following regression:

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

where an observation is a firm-month, Y_{it} is the outcome of interest for firm i, in month t of quarter q, in industry k, and size decile s. In this specification, α_i is a firm-level fixed effect that captures time-invariant firm unobservables while τ_{kqs} are quarter-by-industry-by-size-decile fixed effects which allow for time trends to differ flexibly in each quarter for an industry-size decile combination.[28] We define industries using the ten mutually

[28]We measure size deciles for storage and computation outcomes by using a firm's computation or storage,

exclusive and exhaustive divisions defined by SIC codes.

Each of our coefficients of interest, β_q, represents the difference in outcomes relative to the quarter before the GDPR came into force. Now, because our specification and sample conditioning only use firm information from *before* the first year of the GDPR, we can examine any potential anticipation effects in coefficients directly before the GDPR.[29] Finally, we restrict our analysis to the sample period from July 2015 to March 2020.[30] The identifying assumption of our empirical strategy is a conditional parallel trends assumption. We take advantage of our large sample and allow time trends in our outcomes to vary flexibly by industry and size in our baseline specification, with 110 distinct bins for each quarter (11 defined industries × 10 pre-GDPR size-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:

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

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) estimates the average difference in Y_{it} between treated and untreated firms in the first and second year after the GDPR came into force (relative to the pre-period difference).

4.2 Results

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

Results on Data Storage 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

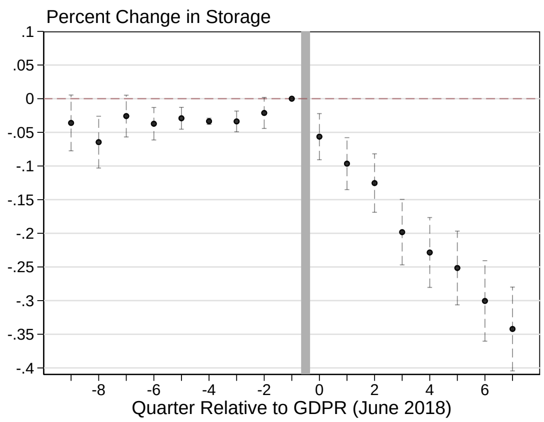
respectively, as measured one year before the GDPR. For data intensity, we use terciles of firm storage interacted with terciles of firm compute, where both outcomes are measured one year before the GDPR.

[29] This specifically refers to relative quarters -1, -2, and -3.

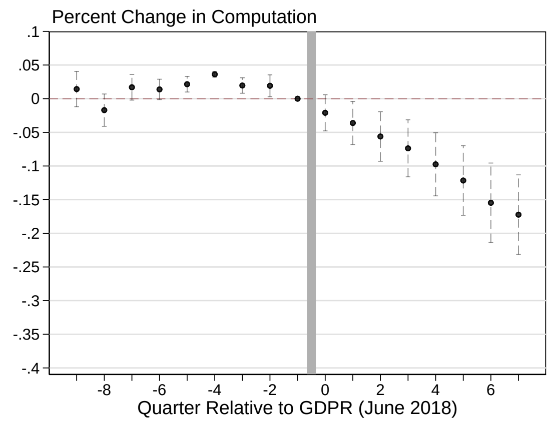
[30] Even though we have data for later periods, we end the sample in March 2020 to rule out the effects of the COVID-19 pandemic. This sample restriction also limits the potential effects of another privacy law, the California Consumer Privacy Act (CCCA), on the US firms in our sample. The CCCA came into effect on January 1, 2020, and applies to businesses that collect the personal data of California residents.

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

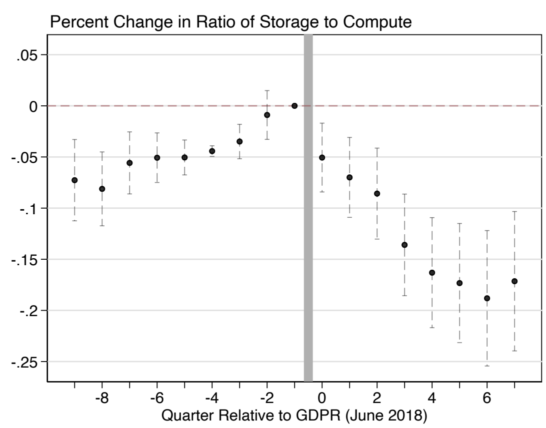
(a) Effect on Storage



(b) Effect on Compute



(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. Gray bars represent the 95% confidence intervals, and standard errors are clustered at the firm level. Sample sizes are presented in Table 3.

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. (2022) in the case of a large website.

As previously discussed in Section 2.1, the decline in data storage 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 storage while the long-run effect doubles to around 26%.[31] This table also shows that our results are robust to the inclusion or exclusion of the flexible time trends by industry and size-decile fixed effects.

Results on Computation Turning towards computation, we first note that there is no clear theoretical prediction for how the GDPR should affect firm computation decisions. GDPR's primary goal is to protect individual data, with little direct impact on computing. Therefore, the effect of the GDPR on computation likely depends on the elasticity of substitution between compute and data and the intensity of these inputs in the production function. If storage and computation are strong substitutes, firms can respond to increases in data costs by substituting away from data toward computation. This would increase total computation. On the other hand, if data and compute are strong 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 decrease their computation relative to US firms after the introduction of GDPR. However, the effect on computation is smaller than what we observe for data storage, with only a 15% decline two years after GDPR. Similar to the results on data, we find no evidence of significant differential pre-GDPR trends in the US and EU.

The results on computation are also important because they indicate that firms do not simply eliminate (or stop accumulating) data they do not use. One potential explanation for our data results could be that, before GDPR, firms stored data that they never utilized and deleted it to comply with GDPR. Our findings suggest that this hypothesis is unlikely to hold because of the substantial reduction in computation.[32]

[31] Importantly, firms are not necessarily deleting data, as our identification strategy relies on comparing EU and US firms within the same industry and size-decile group. Data storage for EU and US firms could be increasing but at different rates.

[32] This hypothesis appears unlikely also because cloud computing incurs a marginal cost for storing data, even

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

	(1)	(2)	(3)	(4)
	Panel A. Dependent variable: Log of Storage			
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 impact 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 compute terciles when measured in the period. Standard errors are clustered at the firm level.

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 computing to storage 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.[33]

Panel (c) of Figure 3 shows that firm data intensity decreases 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 computing are likely complements in production, which we revisit using a production framework in Section 5.

Robustness of Results There are several potential threats to our identification strategy. In Appendix D, we go through the most critical threats to identification and show evidence suggesting that these threats are not driving our results. We summarize the main exercises below, and we leave the additional exercises (such as alternative sample definitions and alternative empirical specifications) and details in Appendix D.

The most salient identification threat is that we observe only one cloud service provider (Appendix D.1). What we observe as declines in cloud usage could simply be firms substituting usage towards other providers. We first show that our results are similar when we restrict our sample to firms that only use our cloud provider (Table OA-2 and Figure OA-6). Therefore, it is unlikely that the declines we observe are simply driven by substitution in usage to other providers. Second, we show that results are unlikely to be driven by firms shifting to traditional (i.e., in-house) IT services. To do so, we show that our empirical exercise yields similar results for the start-up firms in our sample, which are unlikely to have or use traditional IT (Table OA-4 and Figure OA-8).

Another natural explanation for our results is the possibility of differential price trends in the EU and the US (Appendix D.2). If cloud computing providers increased their prices in the EU relative to the US around the time of the GDPR (perhaps to cover GDPR compliance costs, for example), we could see a decline in storage and computation even without the GDPR having direct effects on firms. To check this hypothesis, we use the paid prices for cloud storage as a dependent variable. Appendix Figure OA-9 shows that prices did not change differentially in the EU and the US. Cloud prices have been generally trending downwards, but not in a differential manner between the EU and the US.

if it remains unused. Additionally, in Section 5, we find that firms are responsive to changes in cloud prices.

[33] To increase power, we slightly modify equations (1) and (2). Instead of using size deciles, we construct terciles for storage and compute and interact both to have nine (instead of ten) size bins.

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 (Appendix D.3). We might expect firms with active website use—which we proxy for through the usage of cloud-based web services in our cloud provider—to be more affected by the policy than those without. Table OA-5 shows larger treatment effects among firms that used web services in storage and computation. However, we find that the storage and computing 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 storage 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 GDPR on data and computation vary across four mutually exclusive and exhaustive industry groups: software firms, non-software service firms, manufacturing firms, and all other industries. Table 4 shows our estimates of the short- and long-run impacts of the GDPR when we estimate equation (2) across different industry groups.[34] One striking result is that the direction of our primary findings—declines in storage and computation and decreases in data intensity—are the same across all industry groups. Furthermore, there are detectable effects in storage and computation across all industries. This immediately suggests that our results are not being driven by a single industry and that the effects of the GDPR are not simply limited to software firms, but instead affect firms using data across all industries.

Furthermore, we find substantial heterogeneity between industries in the magnitudes of the effects. Panel A shows that the most significant decreases in storage in response to the GDPR come from manufacturing firms (40% in the long run), followed by non-service and non-manufacturing industries (35%), then by software firms (25%), and non-software service firms (18%). Similarly, Panel B shows that for computation, the fall is largest in magnitude for manufacturing (32% in the long run), followed by non-service and non-manufacturing industries (16%), and then service firms (15% for software and 10% for non-software services in the long run).

While it may seem surprising that IT intensive industries like software and non-software service firms seem to be less impacted by the GDPR, this may reflect differences in the ability of firms in a given industry to shift away from data in their production functions or compliance cost. For example, manufacturing firms might simply be able to substitute

[34] 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 GDPR
(Heterogeneous Effects by Industry Classification)

	Baseline (1)	Software Services (2)	Non-Software Services (3)	Manufacturing (4)	Other Industries (5)
	Panel A. Dependent variable: Log of Storage				
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 Compute				
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 non-software service 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.

Table 5: Effect of Strictness
on Short- and Long-Run Effects of GDPR)

	Storage (1)	Compute (2)	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-average 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 above-average enforcement countries. The long-run term repeats the same procedure but uses the long-run post-GDPR period instead. Strictness is measured according to Johnson et al. (2022) using data from European Commission (2008). We continue to define industries 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 compute terciles when measured in the period. Standard errors are clustered at the firm level.

from data to capital and labor more efficiently than other industries or they might have higher compliance costs. Similarly, service firms may be less responsive to the GDPR simply because storage and computation are essential parts of their production processes.

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

By Enforcement Stringency Although GDPR harmonized the regulation surrounding data protection, enforcement was left at the discretion of each country's data protection authority. In practice, enforcement stringency has varied across countries, partly because of the resource availability to each data protection authority (Johnson, 2022). Because of these differences, one can expect firm behavior to depend on the likelihood they get fined and for firms in countries with more lenient regulators to respond less to GDPR. We use a measure of strictness created by Johnson et al. (2022) using data from European Commission (2008) that varies at the country level to assess how firms' responses vary with how strict their regulators are.[35] We collapse this measure above and below the within-sample average strictness and assign each firm their country's regulator's strictness.

[35]This measure assigns a z-score to each country based on the perception of firms within that country about the regulators' strictness. For more information, see Johnson et al. (2022).

We modify Equation (2) by adding two additional coefficients to capture potential heterogeneity by enforcement stringency. First, we add a triple interaction of the short-run post-GDPR coefficient, the EU categorical variable, and a categorical variable indicating firms in above-average enforcement countries. Our second coefficient repeats the same procedure but uses the long-run post-GDPR period instead. Our main coefficients of interest (the triple interactions) measure the short- and long-run differences in Y_{it} for EU firms with above-median strictness relative to those with below-average strictness post-GDPR. Table 5 summarizes the results. The interaction coefficients (although not statistically significant) suggest that countries in above-average strictness countries face larger declines in storage and computation (4 pp. and 4.7 pp. more than those in below-average strictness countries in the long run, respectively). Data intensity decreases *more* for firms in the above-average strictness countries.

4.4 Discussion

Our results so far suggest that EU firms responded to the GDPR by storing less, computing less, and becoming less data-intensive relative to US firms. These results are important for several reasons. First, we provide direct and large-scale evidence that firms comply with the GDPR by significantly reducing their data and computation. Second, we show that the GDPR distorts firms' input choices by changing the composition of data and computation used in firm production. Third, the results are not driven by a single industry, by a single country, or exclusively by website firms that are affected by cookie consent policy, indicating the far-reaching implications of the GDPR across many industries. Fourth, the heterogeneity in our results across industries provide evidence that the effect of GDPR is likely to differ across firms because some firms rely on data more heavily than others.

Although these findings provide insights into the impact of privacy laws on firm behavior and provide direct evidence, they do not offer a comprehensive understanding of firm-specific economic costs. Such an analysis requires understanding how firms use data in production and the different adjustment margins of firms. 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 estimates its structural parameters. We use our framework to consider both how firms use data and computation in production and how privacy regulations might affect these decisions. One key consequence of the GDPR is that firms' data costs are affected. As data serves as an input in production, any regulatory-induced increase in input costs will inevitably impact firms' input choices. Therefore, we model the GDPR as a gap between the actual cost of data and the perceived cost of data. We focus on estimating the size of this wedge and its implications for firms.

Our framework is designed to be flexible in terms of how data and computation are integrated into firm production. There currently is no standardized framework for how data enters the production function, and there is likely tremendous heterogeneity in how firms use data. For this reason, we model only the relationship between data and computation in firm production rather than modeling a full production function with standard inputs such as labor and capital. We introduce the model below.

5.1 Production Function with Data

Firms produce information by processing data, which requires two inputs: data and computation. We assume the following constant elasticity of substitution (CES) form for the information production function:

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

where C_{it} represents the amount of computation performed by firm i in month t, D_{it} is the amount of data stored by firm i in month t, and ω^c_{it} is compute productivity. The parameter $\sigma = (1/(1-\rho))$ is the elasticity of substitution between data and computing.

Our model includes a firm-specific productivity term, ω^c_{it}, to capture heterogeneity in computing productivity.[36] This choice is motivated by the substantial variation in the data intensity of firms, as reported in Figure 2 of Section 3. This heterogeneity can arise for two reasons. First, there could be inherent production technology differences between firms on how they could use data, making the production of information more data-intensive for some firms than others. Second, even if the production technology is the same, some firms may have higher-quality data or better computation tools (e.g., higher-quality software tools and more skilled engineers) to generate the same amount of information with less data. Our paper is agnostic about the source of ω^c_{it}. However, we believe it is essential to account for such heterogeneity.

We also intentionally refrain from specifying how information is integrated into the production function, as firms can use information in different ways. As a result, our model remains general enough to capture several of the common ways that data has been

[36] The literature typically calls this term "factor-augmenting productivity." We use the term "compute productivity" instead of "compute-augmenting productivity" for the sake of brevity.

modeled as using information, including augmenting overall firm productivity (Jones and Tonetti, 2020), serving as an input in production (Bessen et al., 2022), enhancing labor productivity (Agrawal et al., 2019), and enabling firms to target customers better or forecast demand (Eeckhout and Veldkamp, 2022). These include all of the following cases (omitting subscripts for ease of notation):

$$
\begin{aligned}
Y &= f(K, L)\omega(I) && \text{(productivity increasing)} \\
Y &= f(K, L, I)\omega && \text{(input in production)} \\
Y &= f(K, \omega^L(I) \cdot L)\omega && \text{(labor-augmenting)} \\
R &= p(I) \cdot \big(f(K, L)\omega\big) && \text{(price discrimination)}
\end{aligned}
$$

In these examples, K, L, Y, and R are capital, labor, output, and revenue; ω is Hicks-neutral productivity; ω^L is labor-augmenting productivity; and p is the output price. In each specification, information affects a different part of the production function. Our approach, however, relies on estimating input demand functions using a cost-minimization assumption. We therefore do not need to take a stance on how information enters firm production functions or how firms choose how much information to produce.[37] In our framework, we only need firms to choose data and computation optimally to minimize information production costs.

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, where providers typically follow a pay-as-you-go model, and firms can easily adjust their usage of storage and computation hourly. We also assume that firms are price-takers in the input markets for computation and storage. We again view this assumption as reasonable for cloud computing because cloud providers typically post list prices and firms pay by the hour.[38]

In our model, firms minimize the production cost of producing information by taking input prices as given and optimizing their input choices. We use p^c_{it} and p^d_{it} to denote the input prices for computation and storage, which may vary across firms. We observe both the list prices and the actual prices paid by firms. In theory, all firms should face

[37] Even though this limits some counterfactual analysis we could conduct, we consider it a reasonable trade-off given the large-scale nature of our study, which covers many firms and industries.

[38] All cloud providers offer discounts if firms commit to using cloud resources over a specific period of time. These discounts are called “reserved instance” or “committed use” discounts, depending on the provider. These discounts are typically applied to the list price. A survey of 750 large companies conducted in 2023 suggests that only one-third of companies use these discounts (Flexera, 2023). This number is most likely lower during our sample period and among small firms. Moreover, firms that receive quantity discounts can resell or refund their commitments for a small fee for most major cloud providers. Therefore, we believe that linear prices are good approximations even for these firms.

uniform cloud computing prices since they can access all data centers. However, latency effects and switching costs between data centers may restrict firms' ability to use all data centers, leading to different consideration sets for different firms (and thus differential prices). In addition, potential negotiated discounts may also result in heterogeneous prices. Based on the assumptions of variable storage and computation inputs and short-run cost minimization, we derive the following first-order condition for firms' data and computing choices from the CES production function:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma \log\left(\frac{p_{it}^{d}}{p_{it}^{c}}\right) + \sigma \log(\omega_{it}^{c}), \tag{3}$$

where $\gamma = -\sigma \log(\alpha)$. We provide the complete derivations in Appendix E.1. We also show that we get the same first-order condition if we were to include labor (software engineers) in the information production function in Appendix E.2.

According to this first-order condition, the relationship between input ratio and input prices is governed by the elasticity of substitution between these two inputs. When the price of data (relative to compute) is higher, firms may substitute towards compute, with an intensity of σ. A notable feature of this equation is that the elasticity of substitution between compute and data can be estimated from firms' input demand alone, without observing other inputs or outputs. This property arises from the homotheticity property of the CES production function, commonly used in the literature for estimating the elasticity of substitution (Doraszelski and Jaumandreu, 2018; Raval, 2019; Demirer, 2020).

Although our framework expands upon the production function literature by considering computation and data, it does have some limitations. While we account for variations in data quality across firms using ω_{it}^{c}, we assume that data is homogenous within a single firm. This assumption might be strong since, in reality, firms may have different types of data with varying quality. This limitation would become particularly relevant if, for example, the GDPR affected data composition in firms. To relax this assumption, we would need to include different data types in production, which we do not observe. It is worth noting, however, that the assumption of homogenous inputs within a firm is a common practice in production function research, primarily due to data limitations.

Our approach to modeling data in firm production differs from some recent approaches in the literature. Our framework is a partial equilibrium model where data flexibly enters the production function and therefore cannot speak to some of the important and interesting channels of data production and use proposed by recent literature. For example, in Jones and Tonetti (2020), data is endogenously produced by consumption and then directly contributes to the production of ideas. Farboodi and Veldkamp (2022), similar to our paper, models data as information, but it is used to forecast demand. One important way our approach differs from previous literature, however, is that we recognize that data must be processed to generate useful information, and we therefore include computation as an additional input along with data. As the modeling of data in firm production is an active area of research, we view our framework as complementary to the existing literature.[39]

5.2 The GDPR as a Cost Shock to Data

This section incorporates the effects of the GDPR into a production framework. We model the GDPR as a cost shock to data inputs, as they are the main focus of GDPR regulations. While some aspects of the GDPR do pertain to computation, the first-order effects of the regulation are on data, and the impacts of the regulation on data are significantly larger. Furthermore, computation is less salient to regulators than data, which affects firms' perceived GDPR costs.[40]

The GDPR introduced several changes to the ways that firms handle and store data, resulting in increased costs for data inputs. These costs involve both variable and fixed costs. Fixed costs are one-time expenses that do not vary with the amount of data a firm has, such as hiring a data protection officer, developing a data protection management system, and implementing organizational measures. Since these costs are fixed, they do not affect firms' data and computation decisions. The GDPR also introduced variable costs that scale with the amount of data a firm has. For instance, right-to-forget procedures can be seen as a form of variable cost. The more data a firm collects, the more likely it is to receive requests to delete data. Another example is penalties and increased liability risks. The probability of a data breach likely increases with the amount of data that firms collect, leading to a higher likelihood of penalties and liability. Finally, data security costs can scale with the amount of data that firms collect as well. Appendix B.2 provides more details on how the GDPR affected variable cost.[41]

Given this context, our modeling focuses on the change in variable costs. We make the following assumptions about data costs before and after the implementation of 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 marginal cost of data without the GDPR, and $\tilde{p}^d_{it}$ is the marginal cost of data after accounting for the costs introduced by the GDPR. Therefore, λ_i denotes

[39]See Veldkamp and Chung (2023) for an excellent review of this literature.

[40]If the GDPR's impact on computation costs is non-negligible, our wedge estimate will identify the ratio of data to compute wedges. In this case, our estimate of the wedge introduced will be conservative.

[41]This observation aligns with the fact that larger firms tend to receive more substantial fines.

the wedge between the actual cost of data and the total cost that includes complying with GDPR. We model this wedge as firm-specific because compliance costs will likely 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. We follow the literature and model λ_i as a multiplicative wedge (e.g., Chari et al., 2007; Hsieh and Klenow, 2009).

5.3 Identification of Parameters

Our end goal is to estimate two parameters: the wedge introduced by the GDPR (λ_i) and the elasticity of substitution between computation and data. To illustrate the potential identification problems when estimating λ_i and σ, consider the first-order condition in equation (3) after the GDPR for EU firms:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma \log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \sigma \log(1+\lambda_i) + \sigma \log(\omega_{it}^c). \tag{4}$$

This first-order condition reveals a fundamental challenge for identification: the cost of the GDPR, $\log(1+\lambda_i)$, cannot be separately identified from the firm-specific compute productivity post-GDPR. Intuitively, firms may decrease their data intensity either because their compute productivity has increased or because the GDPR has imposed additional data costs. Without additional information, we cannot distinguish these two cases. Therefore, to identify the GDPR wedge, we need to control for changes in firm-specific computing technology. In order to separately identify these cases, we impose the assumption that computing technology can be decomposed as follows:

$$\log(\omega_{it}^c) = \log(\omega_i^c) + \log(\phi_t^c) + \log(\eta_{it}). \tag{5}$$

Equation (5) specifies that the compute productivity term can be decomposed into a firm-specific component (ω_i^c), an industry-specific time trend (ϕ_t^c), and a mean-zero idiosyncratic component (η_{it}). This decomposition suggests that we need to control for (i) $\log(\omega_i^c)$ to identify firm-specific wedges and (ii) $\log(\phi_t^c)$ to identify the level of wedges 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. In particular, we assume that firm-specific compute technology does not change after the GDPR and that EU industries follow the same compute-technology time-trend as the US firms. These assumptions allow us to control for firm-specific computing technology in the second step, where we estimate the cost of the GDPR as a percentage of the observed data input 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 between computation and data and firm-level compute productivity, we use pre-GDPR data and estimate the following equation:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma_1 \log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma_1 \log(\omega^c_i) + \sigma_1 \log(\phi^c_t) + \sigma_1 \log(\eta_{it}), \tag{6}$$

where σ_1 is the pre-GDPR elasticity of substitution. There are two important considerations 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 and time-varying compute productivity shocks (η_{it}). To address this endogeneity, it is important to understand the factors contributing to the heterogeneity and price changes in cloud computing.

Cloud computing providers display their prices for various cloud computing products on their websites, which typically vary depending on the region or country where the data center is located. These posted prices can be considered exogenous because firms are unlikely to be large enough to affect them. In addition, cost improvements and increased competition have played key roles in price changes in the last decade (Byrne et al., 2018). However, the prices that firms pay may differ from these list prices for two reasons. First, firms may have differential preferences over data center locations.[42] These unobserved preferences may generate endogenous price variation. Second, firms may receive a percentage discount from the listed price based on long-term commitments or bargaining power, as discussed earlier.

These two sources of price heterogeneity can create endogeneity. For instance, firms that experience a high compute productivity shock may be more willing to switch between data centers to take advantage of lower prices, resulting in a correlation between the firm's computation productivity and the prices it faces. In addition, firms with high computation productivity may negotiate higher 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) and estimating the input demand function of firms.

We first introduce the broad intuition behind our instrument. Our shift-share design

[42] For example, firms typically choose data centers closer to their operations to reduce latency.

addresses these two potential sources of endogeneity in prices by leveraging two features of our data. First, because we observe both list prices and negotiated prices, we can use changes in list prices to instrument for the changes in negotiated prices. Changes in list prices for data center locations are plausibly exogenous because no single firm is large enough to affect list prices with their changes in productivity. These changes, however, are still predictive of the prices that firms face because discounts are applied to list prices.[43]

Second, we use the fact that we observe data center choices at a high frequency to construct a measure of exposure to specific data centers for each firm and period. By using historical exposure shares rather than contemporary ones, we leverage the fact that these previous decisions are sunk. However, previous data center choices remain predictive of the data centers that firms will use in the future because of the switching costs associated with moving data between data center locations. Transferring data from one location to another can be time-consuming and expensive, especially for large or complex datasets. As a result, firms' location choices are highly persistent over time.

More formally, the shift-share design combines list prices with variation 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 for each service in a given period are calculated as the share of firm i's usage in a given data center relative to the firm's total demand. This differential exposure gives us the following equation for the instrument:

$$z_{it}^{\{c,d\}} = \sum_{l \in \mathscr{L}} s_{il(t-12)}^{\{c,d\}} p_{lt}^{\{c,d\}} \tag{7}$$

where $s_{il(t-12)}^{\{c,d\}}$ denotes firm i's usage share for data center location l as measured 12 months 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.[44] Our exposure shares are lagged by 12 months because contemporaneous exposure 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 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.[45]

[43]We provide more information about cloud computing pricing in Appendix F.1.

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

[45]One example of a potential threat to identification would be if idiosyncratic compute productivity shocks are strongly correlated over time after accounting for aggregate industry time trends, and this caused firms

We use z^c_{it}/z^d_{it} as an instrument for price ratio p^d_{it}/p^c_{it} and estimate Equation (6) for three EU industries (software, non-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 firm-specific compute productivity (ω^c_i) and production technology parameters before the GDPR. We also estimate Equation (6) for US industries over the entire sample period, as US firms do not experience regulatory distortion either before or after the GDPR. This allows us to recover the industry-specific compute productivity trends, ϕ^c_t 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 wedge generated by the GDPR (λ_i) and the EU post-GDPR elasticity of substitution between computing and storage. In particular, we assume that the cost of data after the GDPR is given by: $\tilde{p}^d_{it} = (1 + \lambda_i)p^d_{it}$, where λ_i reflects the cost of the GDPR. Incorporating this into the firm's input demand, we obtain the following equation:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma_2 \log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma_2 \log(1 + \lambda_i) + \sigma_2 \log(\omega^c_i) + \sigma_2 \log(\phi_t) + \sigma_2 \log(\eta_{it}), \quad (8)$$

where σ_2 is the post-GDPR elasticity of substitution. Here, unlike the pre-GDPR input demand equation, the additional term λ_i affects the ratio of computing to storage. The higher the cost of the GDPR, λ_i, the more likely firms are to substitute away from data toward computation. In order to use this equation for identifying λ_i, we make the following assumptions:

Assumption 1. *Firm-specific compute productivity remains the same after the GDPR.*

We note that this assumption still allows for industry-specific trends in computation due to $\log(\phi_t)$, as we can see from Equation (5). The assumption also does not restrict firms' abilities to respond to the GDPR by changing their compute-to-storage ratio. Rather, it implies that the firm-specific component of the underlying information production technology remains the same.

At this point, it is worth discussing our approach and comparing it to the approaches taken in the literature that estimates wedges. The large literature on misallocation identifies distortions as the difference between the marginal product of an input and its price

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.

(Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). The typical approach in that literature assumes that firms have the same production technology. This assumption is needed 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 some heterogeneity through compute 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 compute technology.

We also differ from this literature in that we do not impose a full production function structure. Instead, we use the demand for two variable inputs—one distorted and one not—to identify the wedge. The underlying idea is that by looking at the ratio of inputs, we can net out the sources of distortions that are common to both inputs, such as market power and adjustment costs, and recover the distortion specific to data input. This identification strategy is similar to the approach used in the literature to identify input market power from the wedge in the ratio between one distorted and one undistorted variable input (Morlacco, 2020; Kirov and Traina, 2021).

Assumption 2. *EU and US industries follow the same time trends in aggregate compute technology post-GDPR.*

This is the second critical assumption necessary for identifying the cost of the GDPR. The identification of wedges requires controlling for aggregate changes in compute productivity. Otherwise, the changes in the computation-to-data ratio of EU firms due to GDPR may be attributed to differential aggregate trends in compute productivity in Europe. Therefore, we use the estimated post-GDPR industry trend from the US firms to control for industry trends in the EU. In particular, the parallel trends we find within industries before the GDPR in our reduced-form results are consistent with this assumption.

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

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma_2 + \sigma_2\Big(\log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \log(\hat{\phi}_t)\Big) + \sigma_2\Big(\log(1+\lambda_i) + \log(\hat{\omega}^c_i)\Big) + \log(\eta_{it}), \quad (9)$$

where $\hat{\omega}^c_i$ denotes estimates of compute productivity using pre-GDPR data and $\hat{\phi}_i$ denotes the estimates of compute productivity trend of the US firms. This equation allows us to estimate our main object of interest (λ_i) along with the post-GDPR elasticity of substitution between computing and data.[46] It is important to note that the elasticity of substitution

[46] Appendix F.5 provides useful intutition behind the identification of λ_i. Roughly speaking, the estimated

Table 6: Elasticity of Substitution Results by Industry

Industry	Software		Services		Manufacturing	
	OLS	IV	OLS	IV	OLS	IV
Elasticity of Substitution (σ)	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 storage and computing (σ) across industries. Estimates are presented for pre-GDPR elasticities for EU firms (σ_1^{EU}). Standard errors are calculated using 100 bootstrap repetitions at the firm level.

parameter is specific to the post-GDPR period. Our specification is therefore flexible enough to allow for and to measure changes in firm production technology post-GDPR.

We estimate this equation using post-GDPR data of EU firms to obtain firm-specific wedges. Because this regression involves generated regressions, the standard errors need to account for first-stage estimations. To address this, we use a bootstrap procedure to estimate standard errors. The bootstrap procedure treats firms as independent observations and re-samples firms with replacement. We present estimates using 100 bootstrap repetitions. We provide the details of the estimation procedure in Appendix F.

6 Production Function Estimation Results

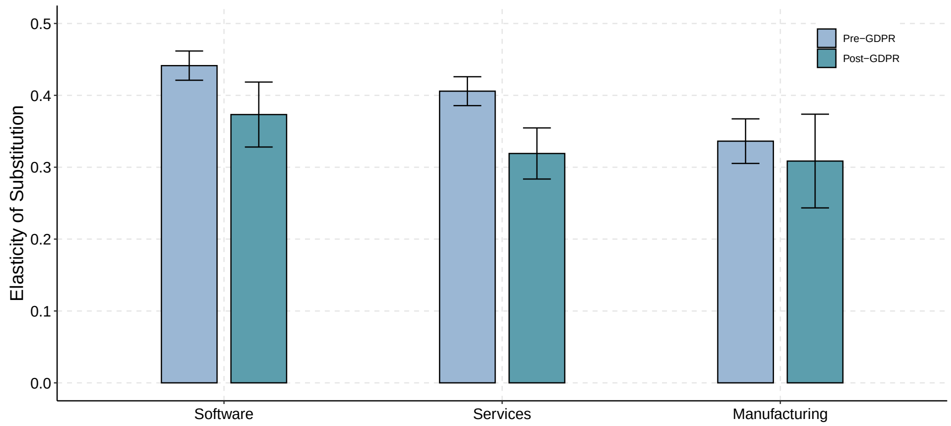
This section provides results on the elasticity of substitution between data and computation, the wedges introduced by the GDPR, and the changes in the cost of information after the GDPR came into force.

6.1 The Elasticity of Substitution Between Data and Computation

We begin by presenting the estimates for the elasticity of substitution using pre-GDPR data. Table 6 presents these elasticities for three industries separately—services, software, and manufacturing—using both OLS and IV estimates. In addition to reporting the estimates for the elasticity of substitution, we also present the first-stage estimates for each industry

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 productivity.

Figure 4: Elasticity of Substitution Between Storage and Computing for EU firms



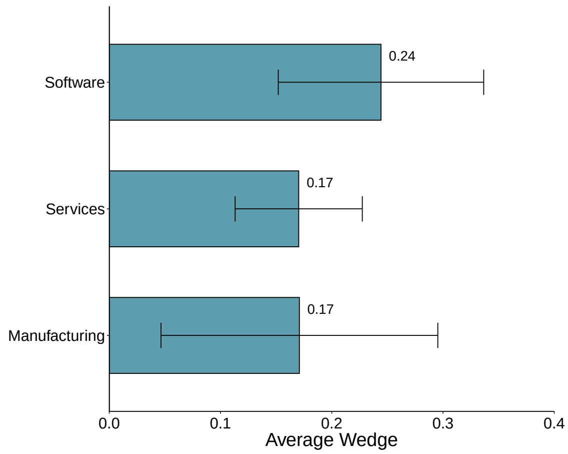
Notes: Figure presents our estimation results of the elasticity of substitution between storage and computing (σ) across industries, and we present separate estimates for the pre- and post-GDPR (σ_1 and σ_2, respectively). Gray bars denote the 95% confidence intervals, and standard errors are calculated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

and 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 high F-statistics, suggesting that our instruments are strongly correlated with the endogenous variables and that we have a robust first stage.

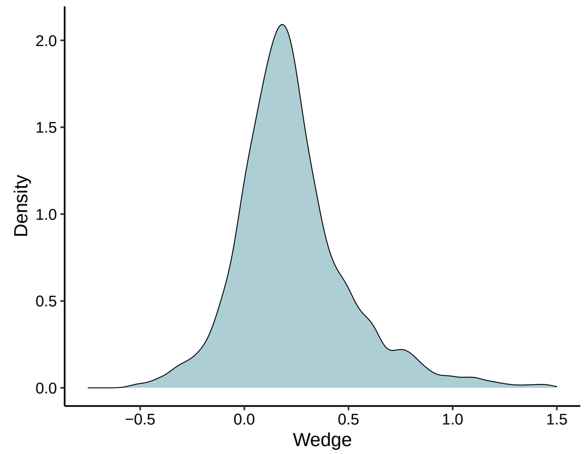
The elasticity of coefficient estimates suggests that data and computation are strong complements in all industries, with an estimated elasticity of substitution ranging from 0.34 to 0.44. The elasticity of substitution is highest in the services industry, suggesting that firms in the services industry can more easily substitute between data and computation. Overall, the complementarity between data and computation is consistent with our reduced-form evidence presented in Section 4, which suggested that firms reduced not only data but also computation in response to the GDPR. Finally, comparing our OLS and IV estimates indicates that using OLS leads to an upward bias in the elasticity of substitution. Thus, as we might expect, the correlation between firms' compute productivity and data-to-compute price ratios is positive; firms with higher compute productivity are more likely to search for lower prices and negotiate higher discounts.

We also investigate how the elasticity of substitution parameters changed after the GDPR, and particularly whether the GDPR led to a change in production technology. Figure 4 reports the elasticity of substitution estimates separately before and after the GDPR for EU firms. While the results suggest a slight decline in the elasticity of substitution in all

Figure 5: Wedge Estimates



(a) Average Wedge by Industry



(b) Wedge Distribution

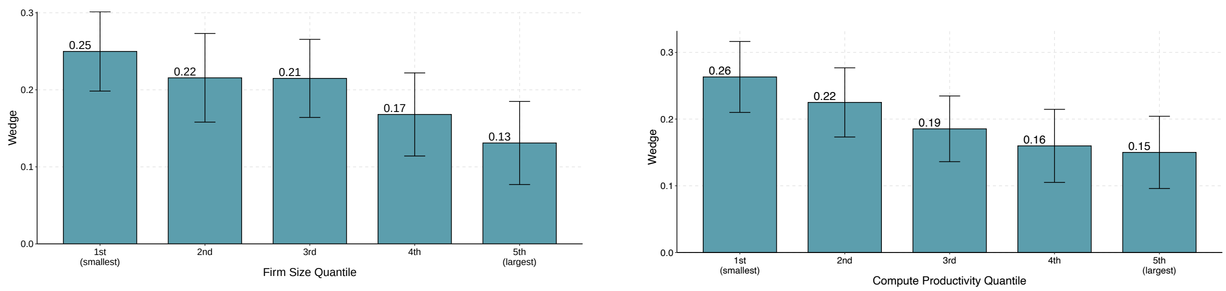
Notes: This figure presents our estimation results for the wedge induced by the GDPR (λ_i). Panel (a) presents the average estimated wedge for firms within each industry. Panel (b) presents the full distribution of estimated wedges. Gray bars denote the 95% confidence intervals, and standard errors are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

industries, we conclude that the GDPR did not lead to a large change in how firms process data to generate information.[47]

Although we are not aware of any previous estimates of the elasticity of substitution between data and computation, it is still informative to compare these estimates with the estimated substitutability between other inputs. The literature has mostly focused on estimating the elasticity of substitution between capital and labor. While estimates vary, evidence with plant-level data suggests values in the range of 0.50 - 0.70 (Caballero et al., 1995; Chirinko, 2008; Raval, 2019). This indicates that data and computation are less substitutable than traditional inputs. Our elasticity of substitution estimates, by themselves, are an important contribution to the literature, as there is very little empirical evidence on how firms use data despite its growing importance. Importantly, the strong complementarity between data and computation suggests that data itself is not sufficient to produce information; firms need to process data, and this requires large computational resources. Therefore, our results highlight the growing role of computation along with data in the modern firm production function.

[47]In Appendix Figure OA-3, we also report estimates for the elasticity of substitution for US firms for comparison. We find that the elasticity of substitution for US firms is similar to that of EU firms and shows no changes after the implementation of GDPR.

Figure 6: Wedge Heterogeneity by Firm Size, Compute Productivity and IT Intensity



(a) Average Wedge by Firm Size

(b) Average Wedge by Compute Productivity

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 productivity (ω_i^c) quintiles computed using pre-GDPR estimates. Standard errors are calculated using 100 bootstrap repetitions at the firm level.

6.2 The Regulatory Wedge Induced by GDPR

Next, we examine our estimates of the GDPR wedge (λ_i). Panel (a) of Figure 5 displays the average wedge for EU firms across various industries together with the 95% confidence intervals. The findings indicate that the average wedge in all industries is statistically significantly different from zero, implying that the GDPR has raised the cost of data for businesses. The wedge is the highest for software firms at 24%, followed by the non-software service industries at 18%. The larger average wedge for software firms could reflect higher average exposure to the costs of the GDPR among software firms. These average estimates, however, hide substantial firm-level heterogeneity. As shown in panel (b) of Figure 5, there is tremendous 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.[48]

To better understand this heterogeneity and to study the determinants of these regulatory wedges, we look at how firm-level variables are correlated with this wedge. We consider two firm characteristics: (i) firm size, as measured by the number of employees, and (ii) compute productivity, as measured by *pre-GDPR* ω_{it}^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-industry. The results suggest that the distortionary effects of the GDPR are highest for the smallest firms, with a wedge equiv-

[48] A small fraction of our wedge estimates are negative, which could be due to noise in the estimation. We find reassuring that most of our wedge estimates are positive, even though we do not restrict them to be so in our estimation.

alent to a 25% tax, and 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 more resources with which to comply with the GDPR. In panel (b), we report the wedge distribution across quantiles of the compute productivity distribution. There is a strong inverse monotonic relationship between compute productivity and the data cost of the GDPR. As firms become more compute-intensive, the magnitude of the wedge decreases from 26% in the first quantile to 15% in the last quantile.

6.3 Cost of Information

How do the additional data costs resulting from the GDPR affect firms' production costs and input decisions? We use the production function estimates to answer this question. We begin by deriving the effect of these wedges on the cost of producing a given level of information or the "cost of information."[49] Given data and computation prices, the cost of information is given by:

$$CI^{*}(I_{it}, p_{it}, \lambda_i) = I_{it} \left((\omega_{it}^{c})^{\sigma} \left(p_{it}^{c}\right)^{1-\sigma} + \alpha^{\sigma} \left((1+\lambda_i) p_{it}^{d} \right)^{1-\sigma} \right)^{1/(\sigma-1)}. \tag{10}$$

Equation (10) suggests that one of the key parameters governing the increase in the cost of information is the elasticity of substitution (σ). Note that if data and computation were perfect complements, then the cost of information would increase linearly with the data cost. In this extreme case, it would be impossible to substitute towards computation when the price of data increases. In contrast, if computing and storage were perfect substitutes, the effect could be minimal for large price changes because firms could fully substitute away from data.

We use the formula for the cost of information given in Equation (10) to estimate the increase in the cost of information post-GDPR by considering two scenarios: (i) a case in which there was no wedge ($\lambda_i = 0$) and the cost of data was simply the cloud cost p_{it}^{d}, and (ii) the realized case in which the cost for firms included the costs of regulations: $(1+\lambda_i)p_{it}^{d}$. To implement this calculation, we use our estimates of key model parameters, such as each firm's compute technology, their input costs, and the elasticity of substitution. These parameters allow us to estimate the counterfactual information cost with and without the privacy regulation for each firm at a monthly level.

The results for the percentage increases in information costs are reported in Figure 7.

[49] The full derivation of the formula for the cost of information is in Appendix E.3.

Panel (a) shows the average change in the cost of information by industry, plotting the mean along with standard errors. These results suggest that changes in the cost of information were significantly lower than changes in the cost of data. The average increase in the cost of information in the manufacturing industry is 2%, while it is about 4% in software and 3% in the services industry. Similarly, Panel (b) documents that considerable firm-level heterogeneity arises from firm-level heterogeneity in data intensities, wedge sizes, and compute productivities.

To further study the heterogeneity of cost increases across firms, we compute the elasticity of the cost of information with respect to the wedge and obtain:

$$\frac{\mathrm{d}CI^*_{it}}{\mathrm{d}\lambda_i}\frac{\lambda_i}{CI^*_{it}} = \underbrace{s^d_{it}\lambda_i}_{\text{direct effect }(+)} + \underbrace{\left[s^d_{it}\left(\frac{\partial D^*_{it}}{\partial \lambda_i}\frac{\lambda_i}{D^*_{it}}\right) + s^c_{it}\left(\frac{\partial C^*_{it}}{\partial \lambda_i}\frac{\lambda_i}{C^*_{it}}\right)\right]}_{\text{firm re-adjustment }(-)} \tag{11}$$

where s^d_{it} and s^c_{it} are the share of total information expenditures spent on data storage and compute, respectively.[50] Equation (11) separates the "direct effect" of the regulation from the "firm re-adjustment" arising from non-local wedge changes. For the first marginal increases in λ_i, the envelope theorem suggests that the direct effect will dominate. As the wedge grows larger, the re-adjustment effect allows firms to absorb part of the increase in costs by re-optimizing data and storage input demand (depending on the value of σ).

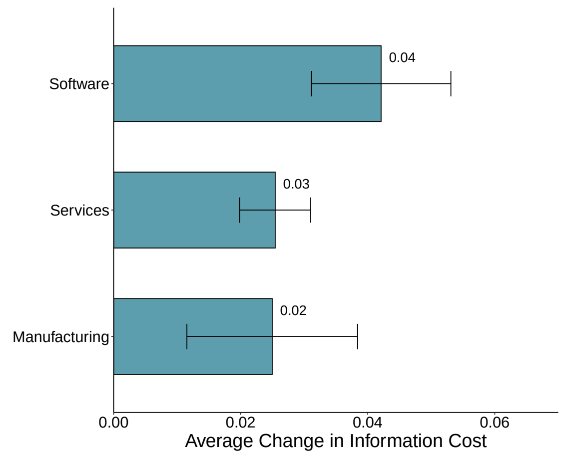
The equally-binned scatterplot in Panel (c) further shows that increases in information costs are highly correlated with the share of total information expenditures. The average increase in the cost of information is 3.7%, and the average cost share of data is indeed around 19%. The figure also captures the heterogeneity in these cost shares, and the firms with the highest data expenditure shares (around 50% of the total expenditures come from data storage) experience an increase in the cost of information of around 13%. Finally, Panel (d) confirms that the direct effects from Equation (11) dominate, as firms are limited in their ability to mitigate the increase in the information cost by substituting data and computation. The figure plots the firm-month distribution of the share of the total change in the cost of information that was mitigated by re-optimization. On average, firms can only absorb 4% of the cost increase by re-optimizing data storage and computation inputs.

To summarize, the GDPR targets the cheaper of the two inputs, storage. An average 20% increase in the cost of data translates to an average 4% increase in the cost of information. The analysis in this section demonstrates the value of our structural approach and estimates of model parameters. By modeling how firms combine data and computa-

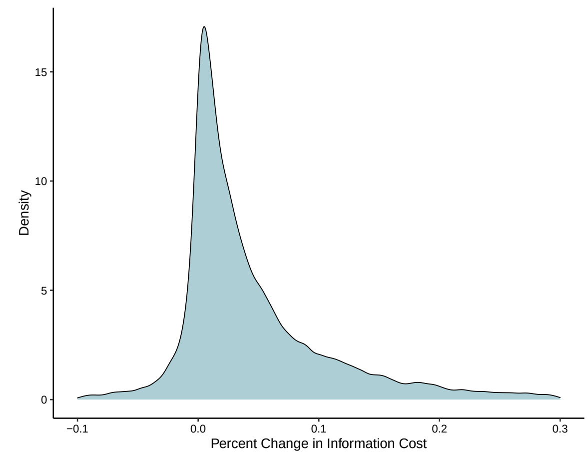
[50] The elasticity can be computed using the firm maximization problem. We compute the (total) derivative with respect to λ_i and multiplying the result by $\left(\lambda_i/CI^*_{it}\right)$. The full derivation is in Appendix E.4.

tion in production, we are able to map the increase in regulatory costs to increases in the production costs of information.

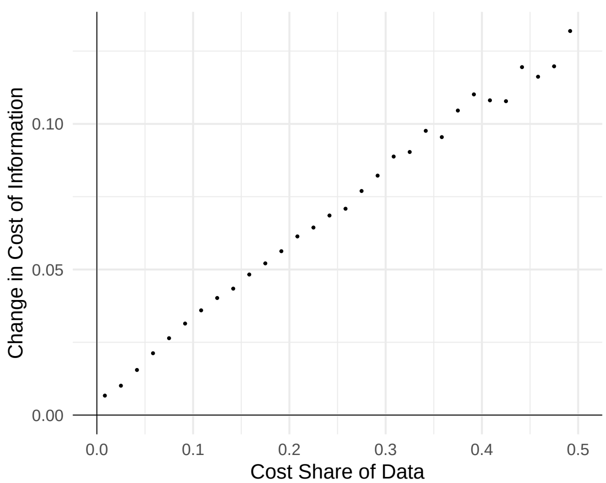
Figure 7: Results on Information Cost



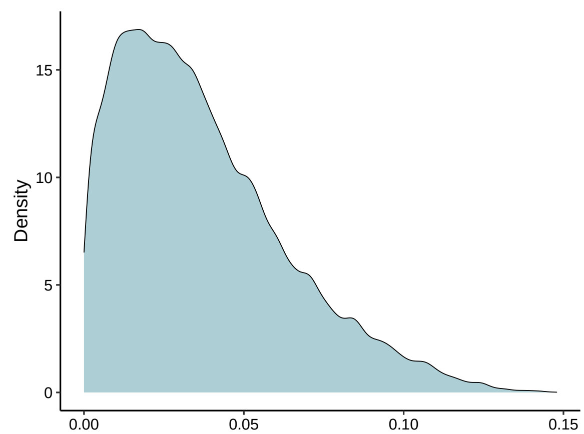
(a) Avg. Change in Info. Cost by Industry



(b) Distribution of Change in Information Cost



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



(d) Firm Re-Adjustment Margin

Notes: Figure presents our estimation results for the change in the cost of information induced by the GDPR. As discussed in the text, we calculate the increase in the cost of information by using Equation (10) to compare the cost of information with our estimated wedge ($\hat{\lambda}_i$) to the cost of information in the counterfactual with no wedge ($\lambda_i = 0$). Panel (a) presents the average estimated increase in the cost of information for firms within each industry. Standard errors are calculated using 100 bootstrap repetitions at the firm level. Panel (b) presents the full distribution of the estimated increase in the cost of information. Panel (c) presents the average estimated increase in the cost of information by the pre-GDPR level of the total expenditures in data. Panel (d) shows our estimates of the "firm re-adjustment" contribution to the total change in the cost of information.

6.4 Production Costs

Finally, to study the impact of the wedges imposed by GDPR on production, we evaluate how our estimated changes in the cost of information translate into changes in production costs. For this goal, one would ideally estimate a production function that captures substitution patterns between information and other inputs. This requires firm-level information on how firms use information and non-data inputs (e.g., capital and labor). In our dataset, however, we do not observe non-data inputs, which precludes us from estimating a full production function.

For the above-stated reasons, we attempt to make progress on this question under some simplifying assumptions and industry-level data. In particular, if the production function is a constant returns to scale Cobb-Douglas, the input elasticities can be measured by their cost shares under the assumption that all inputs are flexible, have common prices, and that firms do not have market power (Foster et al., 2008; Backus, 2020). Using these assumptions, we conduct a simple back-of-the-envelope calculation to estimate how information cost changes translate into production costs.[51]

While these additional assumptions are restrictive, they allow us to provide a benchmark for the effect of the GDPR on production costs using data only on a single input: the cost share of information. Furthermore, this exercise has the advantage of being relatively transparent; it is straightforward to calculate the implied increase in the cost of production for the entire range of our parameters. Denoting the output elasticity of information as γ and the resulting increase in the unit cost function as ζ, we show in Appendix G.1 that under these simple assumptions noted above, an increase in the cost of information (λ) yields the following increase in the unit production cost:

$$\zeta_i = (1 + \lambda_i)^{\gamma_i} - 1. \tag{12}$$

The derivation of Equation (12) captures how firms re-optimize their input mix and substitute away from information in response to the change in information costs. It is also straightforward to calibrate ranges using pairs of firm-level wedges and information shares of expenditure (λ_i, γ_i).

Given the above equation, and since we estimate λ_i, the additional estimates needed are those for the information expenditure shares (γ_i). These estimates are difficult to calculate directly for firms, as there is no standard field for "the cost share of information-related inputs" in most production datasets.[52] Instead, we proxy for information costs by using

[51] Appendix G.1 discusses how the results below change under more flexible production function specifications.
[52] While some researchers have leveraged data from the U.S. Census to track spending on digital technologies,

IT-related expenditures and aim to estimate a range of cost share of information at the industry level.

To estimate the information expenditure shares, we turn to the Aberdeen data set and various industry-level surveys, which we discuss in detail in Appendix G.2. 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. While we might expect each source to suffer from distinct drawbacks, we find that the sources generate remarkably consistent estimates for the information share of expenditure across industries. Appendix Table OA-10 provides the estimates from each source separately, and we take the inter-quartile range from our sources for our back-of-the-envelope calculation.

We present these ranges for ζ from Equation 12 in Table 7. Combining these with the average increases in the cost of information calculated from Section 6.3, we estimate that production costs increase between 0.34% and 0.66% on average for software firms. These average increases are far larger than the ranges we estimate for services and manufacturing firms, which are 0.09-0.15% and 0.05-0.07%, respectively. This difference is primarily driven by the larger information expenditure shares of software firms—the median expenditure share estimate for software is 12.7%, while for manufacturing is 2.7%—combined with the fact that software firms also face the largest average wedges and the resulting increases in the cost of information.

We view the results of our back-of-the-envelope calculation as providing suggestive evidence that the direct impacts of the GDPR that we estimate translated into heterogeneous effects on production costs with non-negligible effects in data and information-intensive industries.

7 Conclusions

In this paper, we examine the impact of the GDPR on firm data input choices. Comparing EU firms affected by the GDPR to similar firms in the US, we document that the GDPR decreased the amount of data used by firms. Firms subject to the GDPR decrease the amount of data stored by 26% and the amount of computation by 15% by the second year after the GDPR, becoming less data-intensive. Our results contribute to the literature documenting the costs of GDPR, complementing the existing literature by focusing on data outcomes that have been rarely studied.

they do not provide relevant industry-level estimates of this statistic that we could use for our estimation (Zolas et al., 2021; McElheran et al., 2023).

Table 7: Effects of GDPR on Production Costs

	Software (1)	Services (2)	Manufacturing (3)
Mean Increase in Information Costs (λ)	0.04	0.03	0.02
Range of Information Expenditure Share (γ)	8.7% - 16.7%	2.9% - 5.0%	2.3% - 3.3%
Resulting Increase in Production Costs (ζ)	0.34% - 0.66%	0.09% - 0.15%	0.05% - 0.07%

Notes: Table presents estimates of equation (12) calibrated with increases in the cost of information estimated in Section 6.3 and a range of information expenditure shares estimated from Aberdeen and other industry surveys for each industry. Column (1) presents these estimates for software firms, which are defined through SIC codes 7370 - 7377. Column (2) presents estimates for non-software service firms. Column (3) presents estimates for manufacturing firms. Appendix G provides more detail about these information expenditure share estimates.

We also map the observed shift in input choices to the production cost of the GDPR using a production function model that we develop and estimate. We are in a privileged position, as we estimate "data usage" as a multi-dimensional object composed of both storage and computing units. We show that storing and computing are complements in production. To our knowledge, these are the first estimates of such a trade-off. Having estimated these results, we then use our model to measure the cost of the GDPR, and we find that the measures that firms had to adopt are equivalent to an increase in the cost of the GDPR of around 20%, with substantial variation across industries. Software industries—that likely find data more useful—are more affected than manufacturing firms, and small firms—that likely find compliance more costly—experience greater distortions in their demand for storage and computation.

There are several potential avenues left to explore in this paper. First, one could leverage additional assumptions about production function and add additional data on output, labor, and capital expenditures in order to recover the full production function and the elasticity of substitution between data, capital, and labor. With such parameters, one could also compute the "data share" in production and measure the extent to which the data share correlates negatively with the labor share, as many other papers have suggested. Second, we left out multinational firms from the analysis, which may have followed different trajectories that are worth studying. Finally, we reiterate that this paper is only a partial analysis of the welfare effects of the GDPR. This paper is completely agnostic to the benefits that consumers derive from the information disclosures provided by the GDPR or the surplus derived from the increased privacy protections that such a law entails. 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.

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

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2022). Too Much Data: Prices and Inefficiencies in Data Markets. *American Economic Journal: Microeconomics 14*(4), 218–56.

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 (2022). The Effect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR. *RAND Journal of Economics* (Forthcoming).

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.

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.

Backus, M. (2020). Why is productivity correlated with competition? *Econometrica 88*(6), 2415–2444.

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

Bergemann, D. and A. Bonatti (2015). Selling Cookies. *American Economic Journal: Microeconomics 7*(3), 259–94.

Bergemann, D. and A. Bonatti (2022). The Economics of Social Data. *RAND Journal of Economics 53*(2), 263–296.

Bergemann, D., A. Bonatti, and A. Smolin (2018). The Design and Price of Information. *American Economic Review 108*(1), 1–48.

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

Bimpikis, K., I. Morgenstern, and D. Saban (2023). Data Tracking under Competition. *Operations Research* (Forthcoming).

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.

Bloom, N. and J. Van Reenen (2007). Measuring and Explaining Management Practices across Firms and Countries. *The Quarterly Journal of Economics 122*(4), 1351–1408.

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*(suppl_1), 299–343.

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).

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, C., C. B. Frey, and G. Presidente (2022). Privacy Regulation and Firm Performance: Estimating the GDPR Effect Globally. *The Oxford Martin Working Paper Series on Technological and Economic Change* (2022-1).

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. (2020). Production Function Estimation with Factor-Augmenting Technology: An Application to Markups. *Working Paper*.

DeStefano, T., R. Kneller, and J. Timmis (2020). Cloud computing and firm growth. *CESifo Working Paper*.

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 Market Power. *NBER Working Paper* (w30022).

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

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.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review 98*(1), 394–425.

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* (Forthcoming).

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. *Rotman School of Management Working Paper* (3112901).

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 (2022). Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR. *Management Science* (Forthcoming).

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. *Denv. L. Rev. 98*, 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.

Kilcioglu, C., J. M. Rao, A. Kannan, and R. P. McAfee (2017). Usage Patterns and the Economics of the Public Cloud. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 83–91.

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 (2021). Labor market power and technological change in us manufacturing. Technical report, Working paper, University of Chicago, Chicago, IL, November 11.

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.

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*.

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*.

McElheran, K., J. F. Li, E. Brynjolfsson, Z. Kroff, E. Dinlersoz, L. S. Foster, and N. Zolas (2023). AI Adoption in America: Who, What, and Where. *NBER Working Paper* (w31788).

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*.

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

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.

Restuccia, D. and R. Rogerson (2017). The Causes and Costs of Misallocation. *Journal of Economic Perspectives 31*(3), 151–74.

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* (Forthcoming).

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 U.S. Firms: Evidence from the Annual Business Survey. *NBER Working Paper* (w28290).

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 . . . . **OA - 2**

B The Impact of GDPR on Firms . . . . **OA - 5**
B.1 GDPR Summary OA - 5
B.2 The Compliance Cost of GDPR OA - 7
B.3 Publicly Available GDPR Fine Data OA - 9

C Data Appendix . . . . **OA - 11**
C.1 Cloud Computing Details OA - 11
C.2 Sample Selection and Cleaning OA - 12
C.3 Aberdeen Sample OA - 13

D Robustness Checks . . . . **OA - 17**
D.1 Substitution to Other Providers OA - 17
D.2 Price Changes OA - 24
D.3 Websites and Cookie Collection OA - 24
D.4 Additional Robustness Exercises OA - 25

E Technical Appendix . . . . **OA - 31**
E.1 First-order Conditions OA - 31
E.2 Including Labor in Information Production Function OA - 32
E.3 Derivation for Cost of Information OA - 32
E.4 Cost of Information Decomposition OA - 33

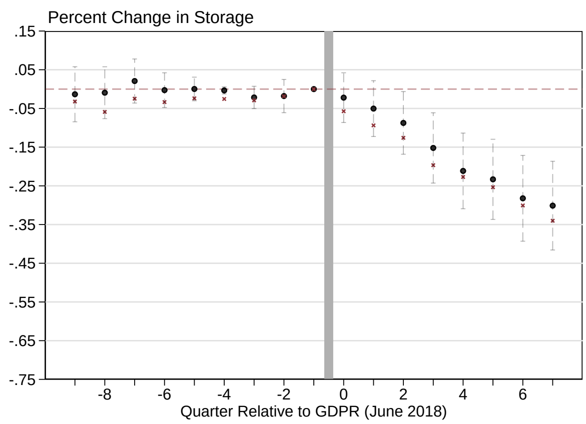
F Model Estimation Details . . . . **OA - 35**
F.1 Cloud Computing Pricing OA - 35
F.2 Price Index Construction OA - 35
F.3 Instrumental Variable Strategy OA - 36
F.4 Estimation Details OA - 37
F.5 Identification Intuition for the Firm-Specific Wedges OA - 38

G Effects on Production Costs . . . . **OA - 41**
G.1 The Effect of Changes in Information Costs on Production Costs OA - 41
G.2 Estimating Key Calibration Parameters OA - 43

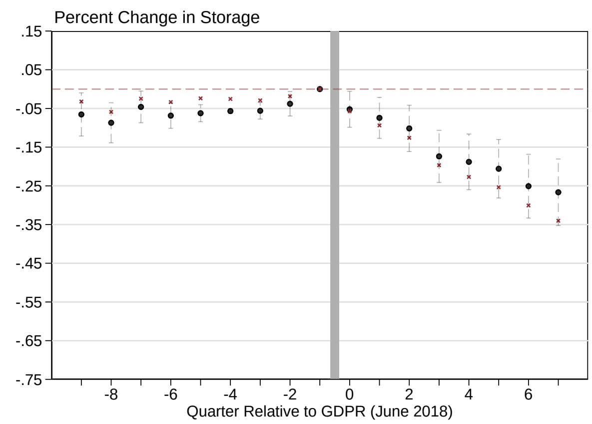
A Additional Exhibits

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

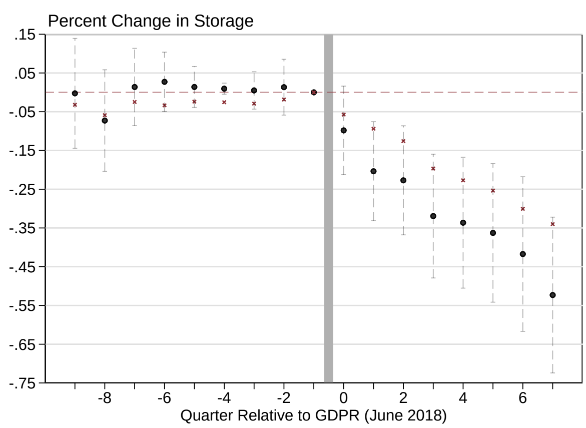
(a) Software Firms



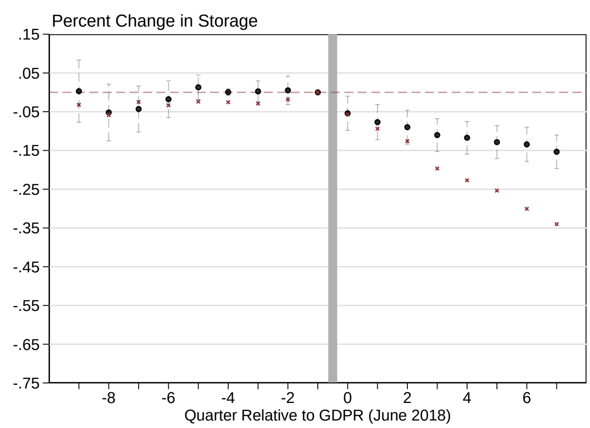
(b) Non-Software Services Firms



(c) Manufacturing Firms



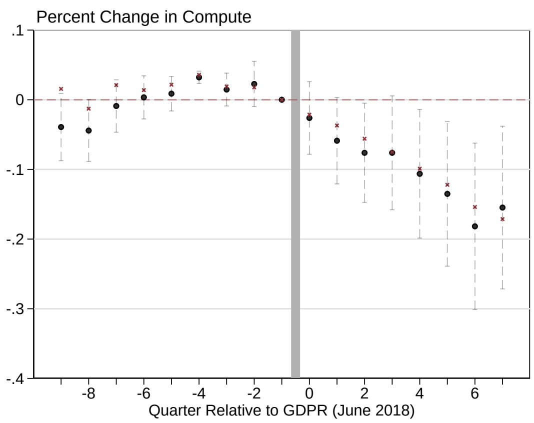
(d) Other Firms



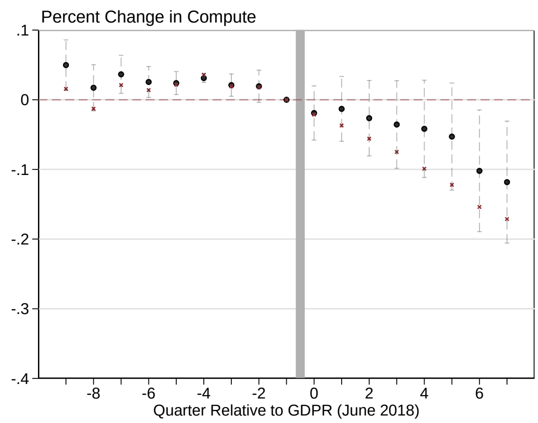
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator, when the outcome is log storage. The coefficient in the quarter before the GDPR's implementation is normalized to zero. Gray bars represent the 95 percent confidence 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 GDPR on Cloud Inputs (Effects on Compute by Industry)

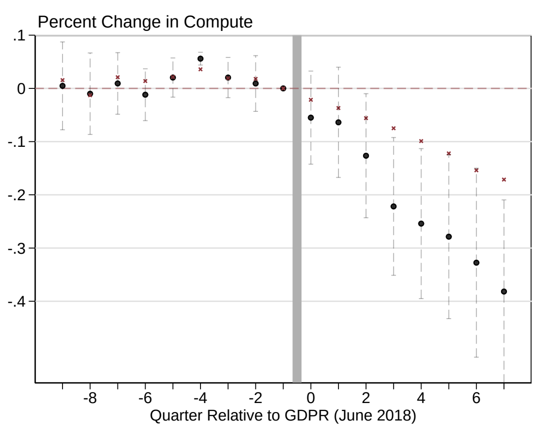
(a) Software Firms



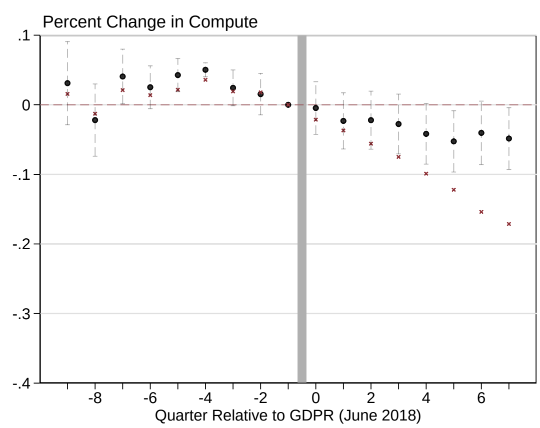
(b) Non-Software Services Firms



(c) Manufacturing Firms

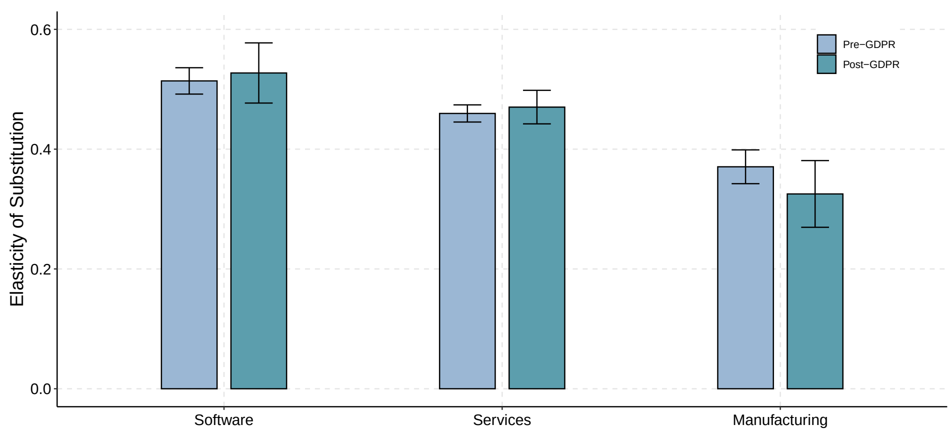


(d) Other Firms



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. Gray 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: Elasticity of Substitution Between Storage and Computing for US Firms



Notes: This table presents our estimation results of the elasticity of substitution between storage and computing (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.

B The Impact of 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).

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). 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 in place technical and organizational measures to ensure the protection of 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 the right to request 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 are able to 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 it. 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 the basic rights and freedoms of individuals. 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 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 cost impact 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. Technology accounts for 12 to 17 percent of total GDPR cost depending on the study. Outside consultants and lawyers accounted for another 19 to 24 percent, depending on the study (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 as a result of 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 insurance covering cyber-risks. Of those insured, 43% had insurance coverage for GDPR fines and penalties.

B.3 Publicly Available GDPR Fine Data

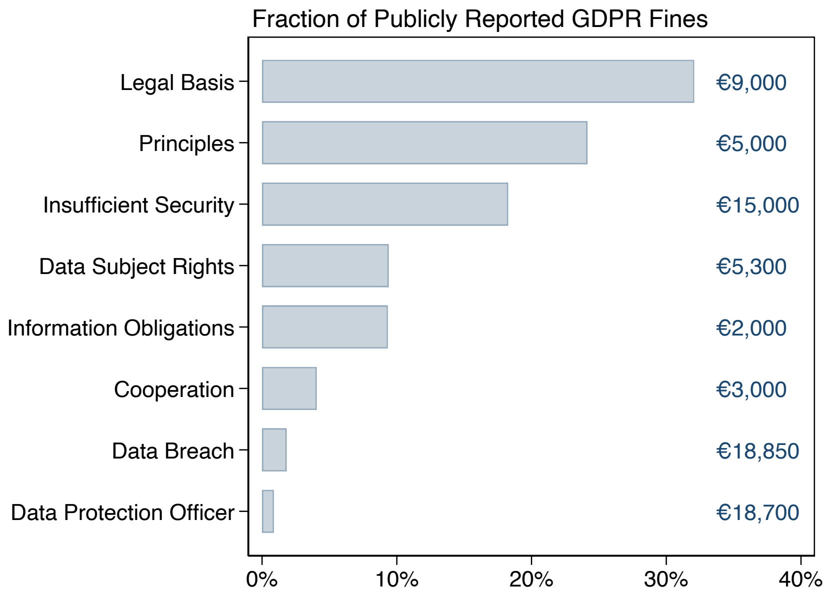
Our primary source of publicly available fine data is a database maintained by CMS Legal Services, a large international law firm that operates in over 40 countries. This data provides an overview of the public fines and penalties that data protection authorities have imposed under the GDPR. Although 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.[53] 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 that it was levied on, the date, and the reason that 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-4, 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

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

Figure OA-4: 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. These categories are described in further detail in Appendix Section B.3. The median fine size by reason is provided in blue text on the right side of the figure.

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 a variety of reasons.

C Data Appendix

C.1 Cloud Computing Details

Cloud computing resources can be categorized into three forms: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS provides storage, computing and networking services on demand. PaaS provides a complete development environment in the cloud, providing low-level infrastructure for development. SaaS provides packaged software services ready to be deployed and used. In this section, we provide details on how firms perform computation and storage in cloud computing.

C.1.1 Computation

Firms that require computation on the cloud typically opt for virtual machines (VMs). VMs are a type of cloud computing “compute" product that allows users to create and manage virtual machines instead of maintaining their own physical hardware.[54] 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 virtual machines that can be customized include memory, storage, networking options, CPU, operating system, and the location of the data center that hosts the VM. 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 characteristics that determines computing performance. We note, however, that this approach does not take into account heterogeneity in other characteristics, such as how much memory and network capability is combined with the number of cores.

The unit of observation is "core hours" which refers to the amount of computing time used by a virtual machine (VM) instance over a given period. The number of core hours used by a VM instance is calculated by multiplying the number of CPU cores by the number of hours the instance is running. For example, if a user runs a VM instance with 4 CPU cores for 10 hours, the total core hours used would be 4 x 10 = 40 core hours. Cloud providers typically use core hours as the relevant measure of VM usage for billing users.

[54] 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.

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 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 that it needs and specify the performance and reliability characteristics that it requires. They would use this disk storage when doing computation in 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 or the amount of storage space used. This is typically measured in gigabytes (GB) or terabytes (TB) and represents a direct measure of how much data firms store, although it does not measure the ways in which 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. In order 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 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. We can see that our we restrict to a larger sample of firms in our baseline sample.

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 provides valuable information on firms' hardware and software investments. They gather this information from various sources. 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 technology adoption information comes only from the latter source. This data also includes sales, the number of employees, industry, and a DUNS number, and these firm characteristics 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 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 Europe and 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, 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 detail 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.[55]

For each firm in the cloud computing dataset, we find the "closest" match in the

[55] 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.

Aberdeen dataset (either by using the parents or the subsidiaries' name). 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 state and the name distance is less than 0.07, or
6. Are in the same country 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.[56] 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 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

[56] 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.

position to offer a partial answer to this question because we possess internal data from one of the largest cloud providers and 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 are 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.

D Robustness Checks

This Appendix goes through the most critical threats to identification. We first 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.

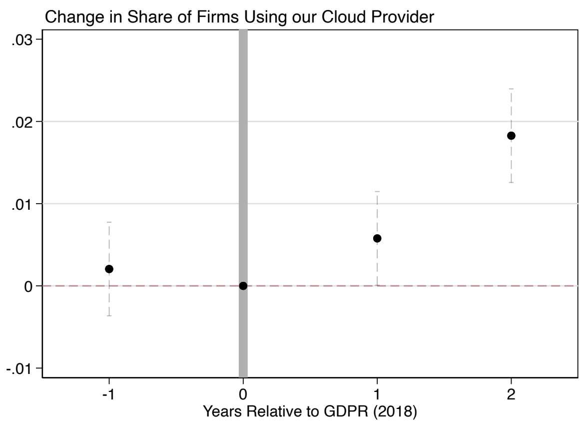
Substitution to Other Cloud Providers "Multi-cloud" usage—where firms get cloud services from multiple cloud computing providers—-is common among firms. Industry surveys suggest that 70 percent of cloud users are multi-cloud. Multi-cloud usage could be a potential issue because we observe usage from 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.

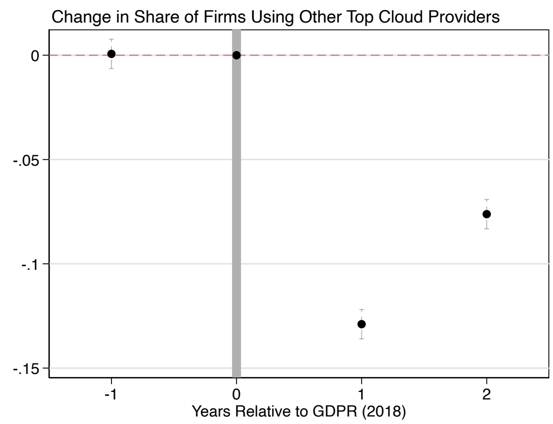
We attempt to address the identification challenge to our storage and computing results with three additional exercises. First, we bring an external dataset, Aberdeen, that provides information on firms' technology adoption and which vendors they get cloud services from. Using this dataset, we look at 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 Appendix Figure OA-5. 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-5: Change in Share of Firms Using Cloud Providers in the EU vs the US

(a) Our Cloud Provider



(b) Other Cloud Providers



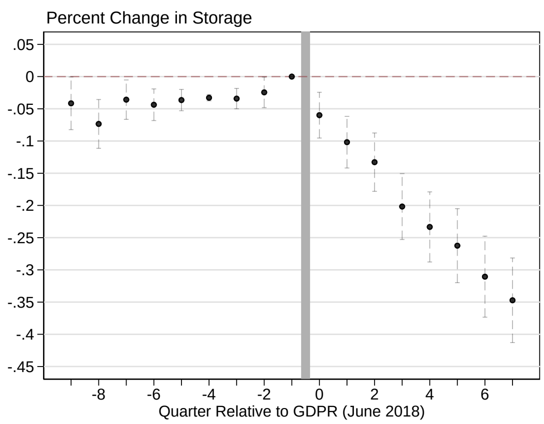
Notes: Figure presents estimates of the difference in the share of firms who use different cloud providers in the EU vs the US. The data source is Aberdeen (formerly known as Harte Hanks). The dependent variable on the left panel is equal to one if a firm uses the cloud provider that we study in this 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. Appendix Table OA-2 and Appendix Figure OA-6 present our estimates using this sample, which are 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.

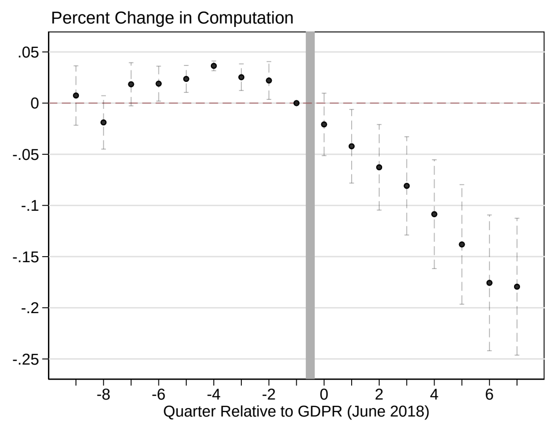
Finally, as discussed in Appendix B.1, the GDPR is likely to make multi-cloud usage more difficult. Thus, switching between cloud providers is more likely to occur on the *extensive* margin rather than the *intensive* margin. Thus, any cloud usage declines in a sample of firms that continuously use our provider are unlikely to be driven by switching between cloud providers. Appendix 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 impacts of the GDPR are quite similar to our baseline estimates. In particular, they are consistent with our findings

Figure OA-6: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Excluding Multi-Cloud Firms)

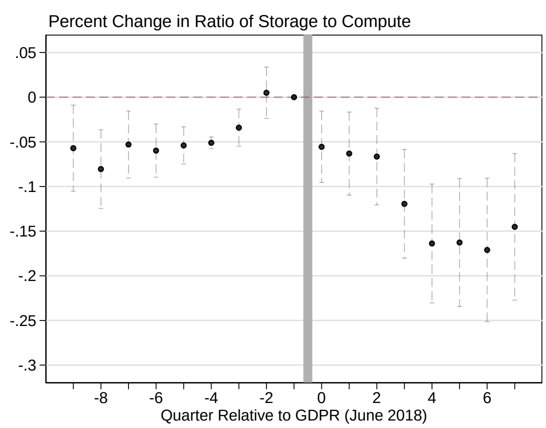
(a) Storage



(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. Gray 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 GDPR (Excluding Multi-Cloud Firms)

	Storage (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.

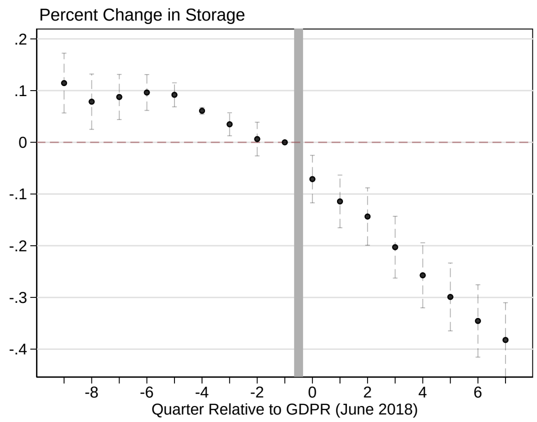
of a large decrease in both compute and storage alongside a decrease in data intensity. Thus, the results from our balanced panel in Appendix Table OA-3 and Appendix Figure OA-7 suggest that the declines in computation and storage we observe are not driven by switching between providers.

Substitution to Traditional IT Next, we consider that firms might use both traditional IT and cloud computing. To the extent that we cannot observe traditional IT usage, declines in cloud computing may reflect re-allocations towards traditional 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 post-GDPR, European firms might have changed allocation between two ITs differently from the US firms.

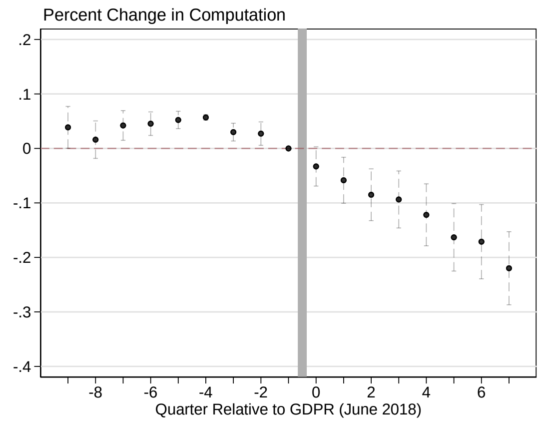
This would invalidate our identification arguments for the effects of compute and storage, though it should not necessarily affect the results on data intensity. To provide a robustness check for this, we focus on start-ups, which are unlikely to be switching to traditional IT. These are young software firms for which the upfront costs of traditional IT make it unlikely for them to switch towards these technologies as they are likely to face larger costs than e.g., more established firms. In Appendix Table OA-4 and Figure OA-8, 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-7: Event Study Estimates of the Effect of GDPR on Cloud Inputs
(Balanced Panel Estimates)

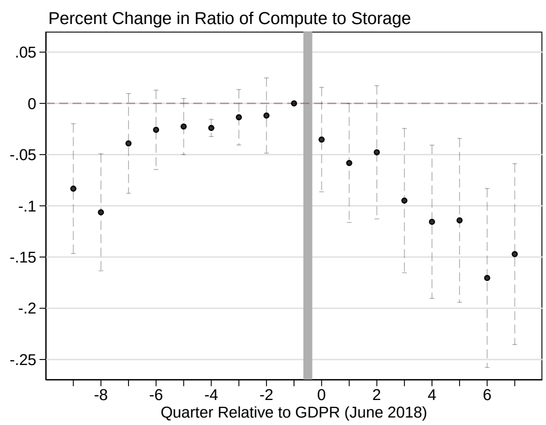
(a) Storage



(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. Gray 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 GDPR
(Balanced Panel Estimates)

	Storage (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 traditional IT.

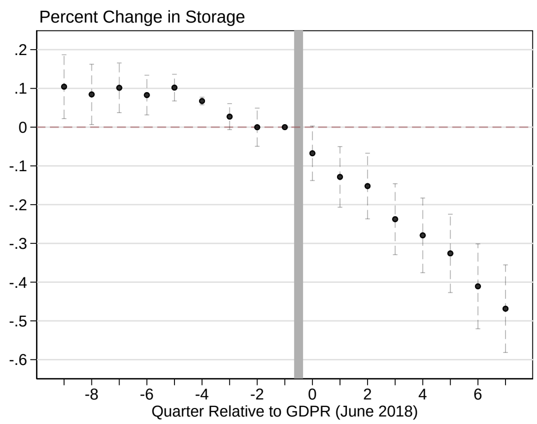
Table OA-4: Short- and Long-Run Effects of GDPR
(Start-Ups Firms)

	Storage (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

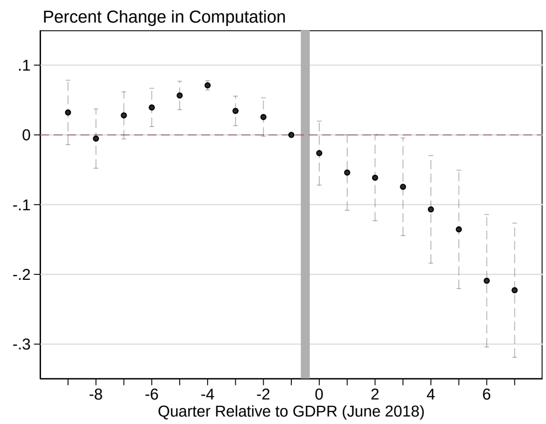
Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. Column (1) estimates the effect on storage. Column (2) estimates the effect on computation. Column (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-8: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Start-Up Firms)

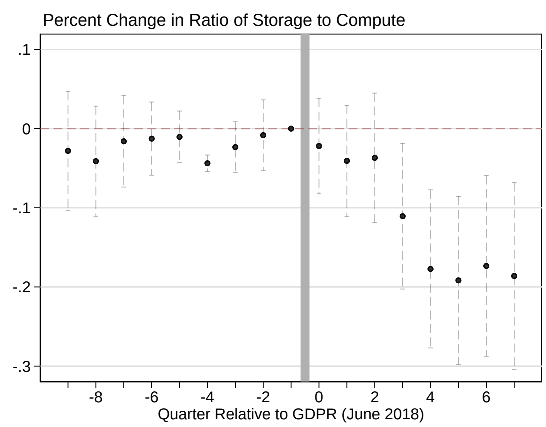
(a) Storage



(b) Compute



(c) Data Intensity



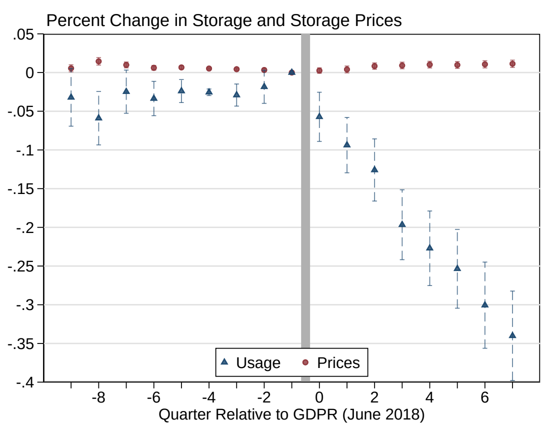
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. The outcome in each subpanel is denoted by the subpanel title. Gray bars 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.

D.2 Price Changes

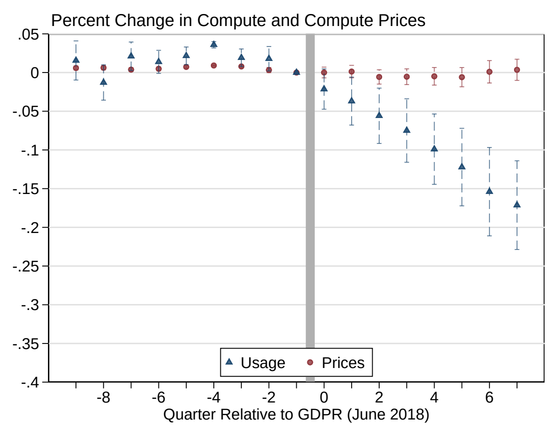
One natural channel through which the GDPR may have affected firms is through price changes in cloud computing. This would suggest our results might capture pricing responses by cloud providers rather than the GDPR's direct impact on firms. For example, if cloud computing providers increase their prices 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. Appendix Figure OA-9 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-9: Event Study Estimates of the Effect of 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. Gray 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. (2022), 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. Appendix 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 only vary by industry, 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 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 Storage			
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 GDPR
(Monthly Specification)

	Storage (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 Appendix 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 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 lograithm 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. Appendix 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 GDPR
(Alternative Pre-GDPR Usage Windows)

	(1)	(2)	(3)
Storage:			
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. Appendix Table OA-9 below presents our estimates using this alternative sample.

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

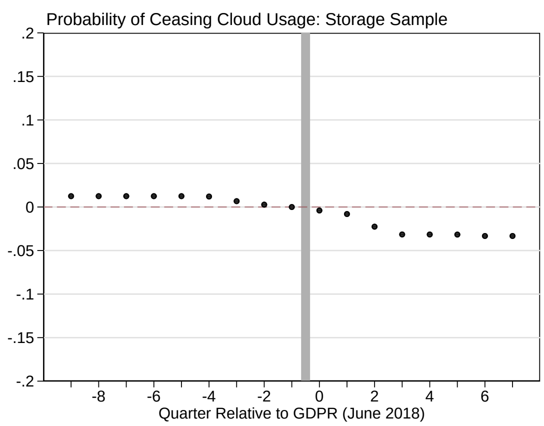
	Storage (1)	Compute (2)	Data Intensity (3)
Short-Run Effect	-0.073	-0.059	-0.063
	(0.013)	(0.013)	(0.015)
Long-Run Effect	-0.151	-0.113	-0.117
	(0.018)	(0.020)	(0.022)
Observations	2,224,810	1,097,922	756,996
US Firms	34,876	18,037	10,807
EU Firms	31,622	15,004	10,299

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.

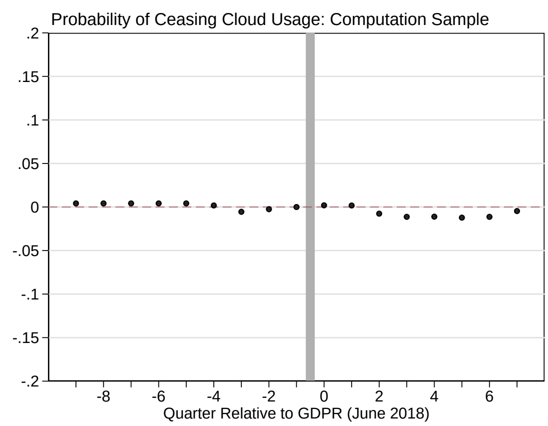
Extensive Margin Although Appendix Table OA-3 suggests that our baseline estimates are similar when we use a balanced panel of firms, we also directly examine whether the GDPR caused differential attrition between firms in the European Union and the United States. We study this using the following same specification but replacing the outcome variable with an indicator for whether the firm has exited our sample. We present these results in Appendix Figure OA-10.

Figure OA-10: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Differential Attrition)

(a) Storage 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. Gray 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 for 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

Assume that firms produce according to the following production function:

$$y_{it} = f(X_{it}, I_{it}, \omega_{it}),$$

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

$$I_{it} = \left(\omega_{it}^{c}(C_{it})^{\rho} + \alpha D_{it}^{\rho}\right)^{1/\rho}.$$

Without loss of generality, we can normalize $\alpha = 1$ due to the homotheticity of the CES production function: $\left(\omega_{it}^{c}(C_{it})^{\rho} + \alpha D_{it}^{\rho}\right)^{1/\rho} = \alpha^{\rho}\left(\omega_{it}^{c}/\alpha(C_{it})^{\rho} + D_{it}^{\rho}\right)^{1/\rho}$.

We assume that firms choose variable inputs to minimize the cost of production taking prices as given, a necessary condition for profit maximization. We also assume that firms take productivity ω_{it}^{c} as given which follows an exogenous process. This cost minimization problem can be written as:

$$\min_{C_{it}, D_{it}} \quad p_{it}^{c} C_{it} + p_{it}^{d} D_{it} + p_{it}^{x} X_{it}^{v} \qquad \text{s.t.} \quad f(X_{it}, I_{it}, \omega_{it}) \geq \bar{Y}_{it},$$

where $\bar{Y}_{it}$ is the target level of production and X_{it}^{v} denotes variable inputs. The FOCs with respect to C_{it} and D_{it} can be written as:

$$\begin{aligned} \lambda_{it} f_2(X_{it}, I_{it}, \omega_{it})\left(\omega_{it}^{c}(C_{it})^{\rho} + D_{it}^{\rho}\right)^{1/(\rho-1)} \rho C_{it}^{(\rho-1)} w_{it}^{c} &= p_{it}^{c} \\ \lambda_{it} f_2(X_{it}, I_{it}, \omega_{it})\left(\omega_{it}^{c}(C_{it})^{\rho} + D_{it}^{\rho}\right)^{1/(\rho-1)} \rho D_{it}^{(\rho-1)} &= p_{it}^{d} \end{aligned}$$

where λ_{it} is the Lagrange multiplier. Taking the ratio of the two FOCs, we obtain:

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

Taking the logarithm and rearranging the terms yields:

$$(1-\rho)\log\left(\frac{C_{it}}{D_{it}}\right) - \log(\omega_{it}^c) = \log\left(\frac{p_{it}^d}{p_{it}^c}\right) \tag{13}$$

By using $\sigma = 1/(1-\rho)$, we can obtain Equation (3) as presented in the main text

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \sigma\log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \sigma\log(\omega_{it}^c). \tag{14}$$

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:

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

Taking the first-order conditions with respect to C_{it} and D_{it}, we obtain:

$$\lambda_{it} f_2(X_{it}, I_{it}, \omega_{it})\left(\left(\omega_{it}^c(C_{it})^\rho + D_{it}^\rho\right)^{v/\rho} + \alpha_L L_{it}^v\right)^{1/v-1}\left(\omega_{it}^c(C_{it})^\rho + D_{it}^\rho\right)^{v/(\rho-1)}\rho C_{it}^{(\rho-1)} w_{it}^c = p_{it}^c$$
$$\lambda_{it} f_2(X_{it}, I_{it}, \omega_{it})\left(\left(\omega_{it}^c(C_{it})^\rho + D_{it}^\rho\right)^{v/\rho} + \alpha_L L_{it}^v\right)^{1/v-1}\left(\omega_{it}^c(C_{it})^\rho + D_{it}^\rho\right)^{v/(\rho-1)}\rho D_{it}^{(\rho-1)} = p_{it}^d$$

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

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

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 nested, the ratio of FOCs would involve labor and our equivalence result would break down.

E.3 Derivation for Cost of Information

In this section, we derive the formula for the cost of information given by Equation (10). To ease notation, we drop the subscript and use p_c and p_d to denote the price of computation and data, respectively. We also use ω in place of ω^c. From the first-order conditions, we obtain:

$$D^{1-\rho} = \frac{p_c}{p_d}\frac{1}{\omega}C^{1-\rho}, \tag{15}$$

which yields:

$$p_d^{\rho/(\rho-1)} C^{\rho} \omega^{\rho/(\rho-1)} = p_c^{\rho/(\rho-1)} D^{\rho}.$$

Adding $p_c^{\rho/(\rho-1)} \omega C^{\rho}$ to both sides of Equation (15) and simplifying yields:

$$Cp_c\big(p_c^{\rho/(\rho-1)}\omega + \omega^{\rho/(\rho-1)} p_d^{\rho/(\rho-1)}\big)^{1/\rho} = p_c^{\rho/(\rho-1)}\big(D^{\rho} + \omega C^{\rho}\big)^{1/\rho}. \tag{16}$$

Similarly, adding $\omega^{1/(\rho-1)} p_d^{\rho/(\rho-1)} D^{\rho}$ to Equation (15) and simplifying yields:

$$Dp_d\big(p_c^{\rho/(\rho-1)}\omega + \omega^{\rho/(\rho-1)} p_d^{\rho/(\rho-1)}\big)^{1/\rho} = \omega^{1/(\rho-1)} p_d^{\rho/(\rho-1)}\big(D^{\rho} + \omega C^{\rho}\big)^{1/\rho}. \tag{17}$$

Adding Equations (16) and (17) and using $I = \big(D^{\rho} + \omega C^{\rho}\big)^{1/\rho}$, we arrive at:

$$\big(Dp_d + Cp_c\big)\omega^{1/\rho} = I\big(\omega^{1/(\rho-1)} p_d^{\rho/(\rho-1)} + p_c^{\rho/(\rho-1)}\big)^{(\rho-1)/\rho}.$$

To derive the cost of information, we need to express the sum $\big(Dp_d + Cp_c\big)$ as a function of I and prices. We do this by isolating the sum on one side of the equation:

$$\begin{aligned}\big(Dp_d + Cp_c\big) &= I\big(p_d^{\rho/(\rho-1)} + \omega^{1/1-\rho} p_c^{\rho/(\rho-1)}\big)^{(\rho-1)/\rho} \\ &= I\left((\omega)^{\sigma}\left(\frac{1}{p_c}\right)^{\sigma-1} + \left(\frac{1}{p_d}\right)^{\sigma-1}\right)^{1/(\sigma-1)}.\end{aligned}$$

Finally, using $\sigma = 1/(1-\rho)$, we arrive at the desired cost function equation.

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

E.4 Cost of Information Decomposition

In this section, we derive the formula for the decomposition of the cost of information given by Equation (11). 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 arguments of the cost-minimizing function. 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:

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

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 Equation (11):

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

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 subsection presents an overview of pricing in cloud computing.

Cloud computing providers typically consider a variety of 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 percent, database prices fell at an average annual rate of more than 11 percent, and storage disk prices fell at an annual rate of more than 17 percent. 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 percent, 22 percent, and about 25 percent for computation, database, and storage, respectively, between 2014 and 2016

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 the discounts negotiated by firms or variation 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 2 percentiles. We also construct the storage-to-compute 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:

$$r_{lt}^{j} = \frac{\sum_i p_{il(t+1)}^{j} q_{ilt}^{j}}{\sum_i p_{ilt}^{j} q_{ilt}^{j}}$$

where $j \in \{c, d\}$ denoting computation and storage, q_{ilt}^{j} 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 $p_{lt}^{j} = \prod_{1 \le j \le t} r_{lj}^{j}$, 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.[57] Even though the specific costs and risks of data migration will depend on the migration plan

[57] See https://aws.amazon.com/blogs/architecture/overview-of-data-transfer-costs-for-common-architectures/, https://azure.microsoft.com/en-us/pricing/details/bandwidth/, and https://cloud.google.com/storage-transfer/pricing for data transfer costs for top cloud computing providers.

and the cloud service provider, it is typically considered too costly by industry experts.

We use the persistance in data center location that comes from switching cost 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_{it}^d and p_{it}^c for data and computation, respectively. This price could be endogenous because the firm may negotiate lower prices 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 further purge these shares of endogeneity by taking lags, as contemporary shares may be susceptible to reverse causality. Hence, our instrument for data is given by $z_{it}^d = \sum s_{i(t-12)l}^d p_{lt}^d$ for storage and z_{it}^c for computation calculated similarly. Finally, we use z_{it}^c / z_{it}^d to instrument for p_{it}^c / p_{it}^d in the production function estimation. Since we need the 12 months 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 assumptions that the industry-specific cloud productivity trend in Europe would have followed that of US firms in the absence of 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:

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma_1^{US} \log\left(\frac{p_{it}^d}{p_{it}^c}\right) + \sigma_1^{US} \log(\omega_i^c) + \sigma_1^{US} \log(\phi_t^c) + \sigma_1^{US} \log(\eta_{it}), \tag{18}$$

When estimating this equation, we normalize γ to zero because it is not separately identified from the mean of ω_i^c. We also normalize ϕ_1^c to 1 so that productivity trend is relative to the initial period. Since, by assumption, the US firms have not been exposed to GDPR, this equation identifies the industry-specific compute productivity trends, or $\hat{\phi}_t^c$ in Equation (9). By Assumption (2), the EU industries follow the same trend and we use the estimated $\hat{\phi}_t^c$ for EU firms.[58] Next, we estimate the same equation using EU firms only with pre-GDPR data. This estimation identifies $\hat{\omega}_i^c$ in Equation (9) because there is no distortion

[58] We also estimate Equation (18) using pre- and post-GDPR data for US firms to separately identify the elasticity of substitution before and after the implementation of GDPR.

before GDPR to estimate σ_1^{EU}. We report the associated 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 (9):

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

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 (9) (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 replacement in each industry-continent group and apply the entire estimation procedure.

We use Equation (10) to estimate the change in the cost of information, with results reported in Section 6.3. For the 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 GDPR. Since prices change over time, we calculate this change in information cost at every observed price point and report the distribution at the month-firm level.

To do the decomposition presented in Equation 11, we calculate the cost share of data every period using firm's data input demand and prices. The direct effect is obtained by multiplying the data share 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 above, we calculate this change in information cost at every observed price point 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 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).[59] Subtracting the pre-GDPR first order condition (Equation 6) of the US firm from the EU firm equation in a period $\underline{t}$ before GDPR implies that:

$$\Delta_i \left(\frac{C_{i\underline{t}}}{D_{i\underline{t}}} \right) = \sigma_1 \Delta_i \left(\frac{p^d_{i\underline{t}}}{p^c_{i\underline{t}}} \right) \tag{19}$$

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 8) in a similar fashion as before (focusing on the same two firms) to obtain:

$$\Delta_i \left(\frac{C_{i\bar{t}}}{D_{i\bar{t}}} \right) = \sigma_2 \Delta_i \left(\frac{p^d_{i\bar{t}}}{p^c_{i\bar{t}}} \right) + \sigma_2 \log(1 + \lambda_i) \tag{20}$$

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:

$$\Delta\Delta_{it} \left(\frac{C_{it}}{D_{it}} \right) = \sigma_2 \Delta\Delta_{it} \left(\frac{p^d_{it}}{p^c_{it}} \right) + (\sigma_2 - \sigma_1) \Delta_i \left(\frac{p^d_{it}}{p^c_{it}} \right) + \sigma_2 \log(1 + \lambda_i) \tag{21}$$

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 (21) 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.[60] Given the above equation, one would intuitively expect firms that face larger

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

[60] 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 GDPR. Importantly, we show that industries with larger changes in compute-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 non-software services, and then by software services. Similarly, our average wedge estimates (shown in Figure 5) have the same ordering: manufacturing firms face smaller wedges, followed by non-software services, and finally by software services.

Interestingly, Equations (21) and (20) 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 responses of compute and storage, one would need additional assumptions over 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), which we intentionally refrain from doing in this paper.

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, the spirit of this exercise to provide a back-of-the-envelope calculation 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 for labor and capital and that the cost function is given by:

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

We first consider the 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. Finally, 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.

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:

$$\begin{aligned} L^* &= \alpha\bar{Y} \\ K^* &= \beta\bar{Y} \\ I^* &= \gamma\bar{Y} \end{aligned}$$

In this case, the cost function is therefore linear in prices, and a λ percentage increase in the cost of information causes an $\lambda \cdot s_{it}^I$ 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 λ percentage increase in the cost of information translates into a zero percentage increase in the cost of production.

Cobb-Douglas Production Function

Finally, we consider the effects of a λ 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 λ percentage increase in p^I induces a $\delta = \left[(1+\lambda)^{-\frac{\alpha+\beta}{\gamma+\alpha+\beta}} - 1\right]$ percentage decrease in I^*.[61] Next, we note that first-order conditions imply that a γ share of total firm costs will be spent on information:

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

Using the change in information expenditure resulting from the λ increase in information prices and the δ decrease in I^* derived above, we have that a λ percentage increase in p^I will lead to a ζ percentage increase in production costs, where $\zeta = (1+\lambda)^{\gamma} - 1$.[62]

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

[62]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 $\lambda = 0$.

G.2 Estimating Key Calibration Parameters

We show in the section above that under a Cobb-Douglas production technology assumption, we only need to know a single parameter—γ—to know how an increase in the cost of information translates to production costs. We note that γ represents the information share of expenditure, and we combine various data sources to suggest a reasonable range for this share. We provide all of these estimates in Table OA-10, and we discuss each of these sources in greater detail below.

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 non-software service firms. Column (3) presents estimates for manufacturing firms. Further details on the Aberdeen data and the survey estimates are provided in Appendix G.2.

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, non-software services, and manufacturing. We find that, somewhat unsurprisingly, software firms spend the highest share of their revenue on IT, followed by non-software services and then manufacturing.

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.