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

Mert Demirer [”] Diego Jiménez-Hernández [···] Dean Li [Ÿ] Sida Peng [¶]

October 30, 2023

Abstract

By regulating how rms collect, store, and use data, privacy laws may change the role of data in production and alter rm demand for computation and data storage. We study how rms respond to privacy laws in the context of the EU's General Data Protection Regulation (GDPR) by using seven years of con dential data from one of the world's largest cloud-computing providers. Our di erence-in-di erence estimates indicate that, in response to the GDPR, EU rms decreased data storage by 26% and processing by 15% relative to comparable US rms, 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 nd that data and computation are strong complements in production and that rm responses are consistent with the GDPR representing a 20% increase in the cost of data on average, with smaller rms 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:

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 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 us to draw more generalizable conclusions about rms' data uses, the trade-o is that we non-GDPR countries.[5] Johnson (2022) provides a comprehensive survey of this literature.[6]

While our paper builds on an identi cation strategy similar to some of these GDPR papers, it is di erent in two main aspects. First, because of the richness of our data, we directly study rms' data and computation decisions, a margin that is the key target of regulation. In particular, our data is well-suited for studying rm 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 rms.

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 rms in the economy. Farboodi and Veldkamp (2022) model data as a productivity-enhancing input that helps rms accurately predict future outcomes. We complement this literature by developing and estimating a rm production framework with data, providing empirical estimates on how rms combine data and computation.

Third, our paper is related to the literature on misallocation, which documents large di erences in the e ciency 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 rms' input allocation. We employ a similar identi -cation strategy by modeling regulation as a wedge between the m5(he)-366dl es Kl1 l esestory

2 Institutional Setting

This section rst 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 rms but because of its crucial role in shaping privacy laws. In this section, we describe the key features of this policy and how they a ect rms.

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 rms 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 rms to protect data, in addition to a substantive system of individual rights. This increases the cost of handling data for rms. 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 e ective manner." This risk-based approach makes costs heterogeneous across rms based on the sensitivity of data and rms' risk preferences.

The GDPR applies whenever the rm (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 de ned 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 e ect, leading to variation in its implementation across di erent 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 rms are subject to GDPR.

From the rm perspective, the GDPR primarily increased the cost of collecting and storing data by imposing costly responsibilities on rms. These include keeping a record of processing activities (Article 30), designating a data protection o cer (Article 37), preparing data protection impact assessments (Article 35), implementing appropriate technical and organizational measures for data security (Article 32), providing timely noti cations in case of data breaches (Article 33), executing consumers' requests for data transfer, erasure, or recti cation (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 signi cantly depending on the size and complexity of an organization. There are no o cial statistics, but most survey evidence suggests that complying with the GDPR is costly for rms. 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 rms. The survey evidence indicates that a large percentage of the costs (between one- fth 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 xed and marginal costs. For example, the cost of having a data protection o cer may not scale with data size, so the latter could be considered mostly a xed 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 rm size. As such, it may be more sensible to interpret these kinds of costs as changes to the marginal costs. We provide a detailed classi cation of GDPR costs into these xed and variable cost categories and present corresponding survey evidence in Appendix B.2.

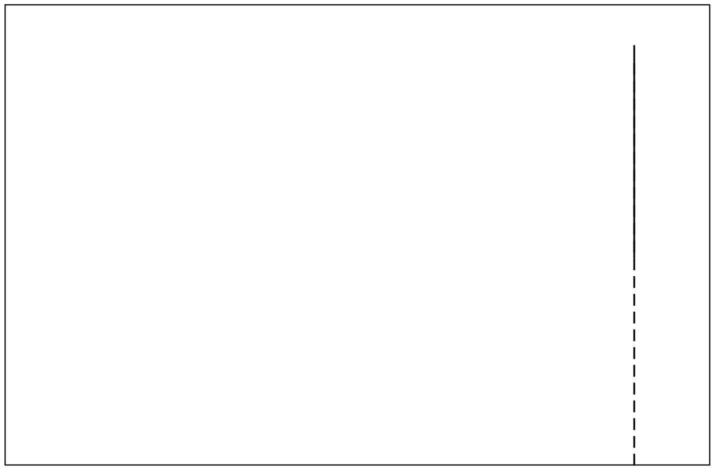
In addition to these direct costs, organizations may also incur indirect costs such as cybersecurity insurance or penalties if they are found to be non-compliant or in the case of data leaks.[10] Non-compliant rms may face nes of up to 4% of an organization's annual globalrevenue or ¿20 million (whichever is greater). We scrape publicly available GDPR

---

9 Contrary to popular belief, consent is not the only appropriate legal basis that rms 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, rms are required to identify which legal basis they are using to process personal data.

10 There are likely additional costs beyond the direct nancial 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 gure presents the distribution of 1,730 publicly available GDPR nes, noting that not all GDPR nes 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 unde ated nominal terms (¿), and ve examples from the data have been highlighted: a restaurant, a jewelry manufacturer, Google, Amazon, and Meta.

ne data (which we describe detail in Appendix B.3) from a database maintained by CMS, an international law rm. [11] In Figure 1, we provide the size distribution of these GDPR nes. [12] We note two key features of these nes. First, the distribution of ne sizes implies that enforcement is not limited to large violations: 25% of the nes 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 rms. 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

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 rm is a start-up or not. [21]

One limitation of our dataset is that it does not allow us to see which speci c data rms are collecting nor the exact ways in which they use the data. This limits our ability to speak to some important questions about how rms speci cally 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 rm 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 rms and examine di erential changes in market share around the GDPR for cloud providers.

3.3 Other Datasets: Firm Characteristics

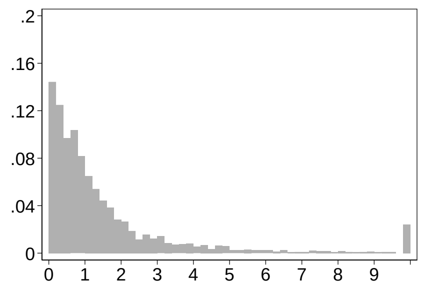
Aberdeen also provides information on other rm characteristics, such as employment and revenue from Duns & Bradstreet. We match our cloud computing data to Aberdeen rms using a matching procedure described in Appendix C.3 based on name, location, domain, and other information. We are able to match close to 60% of our cloud rms to the Aberdeen dataset. We use the employment information in 2018 to de ne rm size. We further augment our data by d0 r in 2a9uiTd (W)65(e)-2ee

Table 2: Summary Statistics

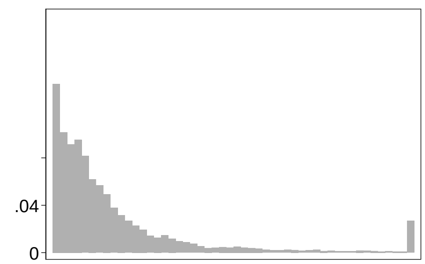
Industry	Number of Firms	Share Compute	Share Storage	Mean Storage	Mean Compute	Mean Data Intensity	Share EU
Services	15,886	36.3%	31.9%	844	628	1.84	40.9%
Software	9,480	17.6%	20.8%	690	670	1.69	59.8%
Manufacturing	3,095	10.5%	11.6%	1,293	986	1.81	54.4%
Retail Trade	2,152	5.2%	5.4%	1,101	917	2.02	46.9%
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:

Figure 2: Histogram of Data Intensity by Industry

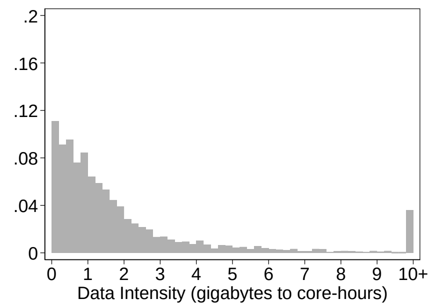


(a) Software Firms



(b) Non-Software Services Firms

(c) Manufacturing Firms



(d) Other Firms

Notes: Figure presents a histogram of data intensity at the rm level, de ned as the ratio of data stored to computation (the ratio of gigabytes to core hours) for each industry, which de ned by SIC codes (with the exception of software rms, which are carved out of the services division). We limit to the sample of rms who have ever used both storage and computation (

4 Event Study Evidence

In this section, we apply an event study design to study the e ect of the GDPR on rms' data storage and computing decisions. We begin by de ning 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 speci cations. Finally, we estimate how the e ects of the GDPR vary across industries in our sample.

4.1 Empirical Strategy

Our empirical strategy aims to identify the causal e ect of the GDPR on rms' computation and data choices. In order to identify a relevant treatment and control group for our strategy, we turn to our classi cations of rm locations from Section 3. Following Table 1, we de ne Case 1 as our treatment group and Case 4 as our control group.

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 $\times$ 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:

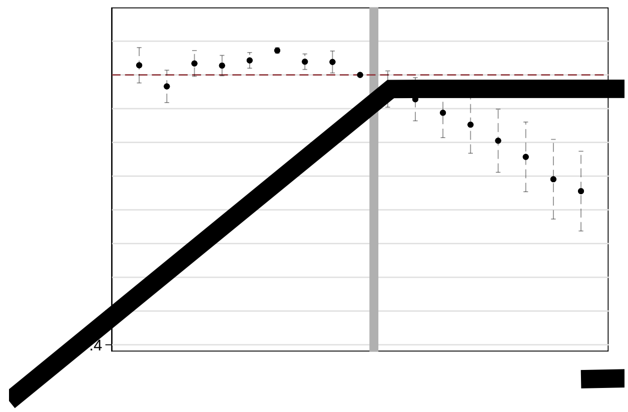
$$y_{it} = \beta_1 1\{EU_i\} \times 1\{t \in \text{Jun/18-May/19}\} + \beta_2 1\{EU_i\} \times 1\{t \in \text{Jun/19-May/20}\} + \alpha_i + \delta_{kqs} + \varepsilon_{it} \quad (2)$$

where the notation of α_i and δ_{kqs} is the same as in equation (1

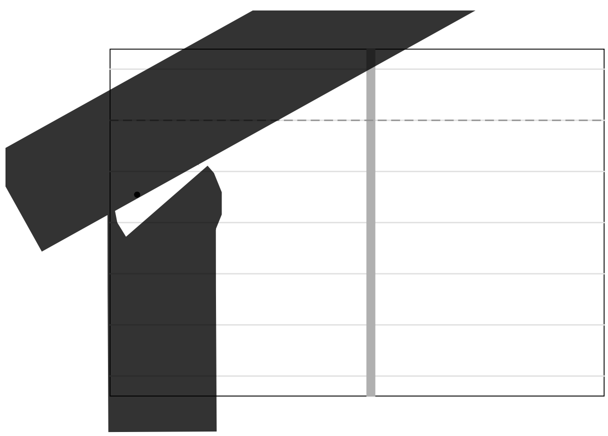
Figure 3: Event Study Estimates of the E ect of GDPR on Cloud Inputs

(a) E ect on Storage

(b) E ect on Compute



(c) E ect on Data Intensity



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

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 ($_1$) and long-run ($_2$) 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 rms became less data-intensive after the GDPR. However, in order to account for potential compositional e ects, we investigate the e ects of the GDPR on data intensity by using the natural logarithm of the ratio of computing to storage as an outcome. We estimate our speci cation on rms 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 rm 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 rms in the EU become less data-intensive post-GDPR (relative to comparable US rms) 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 identi cation strategy. In Appendix D, we go through the most critical threats to identi cation 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 de nitions and alternative empirical speci cations) and details in Appendix D.

The most salient identi cation threat is that we observe only one cloud service provider (Appendix D.1). What we observe as declines in cloud usage could simply be rms substituting usage towards other providers. We rst show that our results are similar when we restrict our sample to rms 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 rms 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 rms 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 di erential 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 e ects on rms. 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 di erentially in the EU and the US. Cloud prices have been generally trending downwards, but not in a di erential manner between the EU and the US.

---

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

33

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.e

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

	Baseline	Software Services	Non-Software Services	Manufacturing	Other Industries
	(1)	(2)	(3)	(4)	(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 $_1$ and $_2$, 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: E ect of Strictness
on Short- and Long-Run E ects of GDPR)

	Storage (1)	Compute (2)	Intensity (3)
Short-Run E ect	-0.028	-0.061	-0.042
	(0.044)	(0.032)	(0.042)
Long-Run E ect	-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 e ect of above-average GDPR strictness. The short-run term captures the triple interaction of the short-run post-GDPR coe cient, the EU categorical variable, and a categorical variable indicating rms 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 de ne industries as the ten divisions classi ed by SIC codes. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we de ne size decile as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the rm level.

from data to capital and labor more e ciently than other industries or they might have higher compliance costs. Similarly, service rms 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 nd 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 storageputation.

We modify Equation (2) by adding two additional coe cients to capture potential heterogeneity by enforcement stringency. First, we add a triple interaction of the short-run post-GDPR coe cient, the EU categorical variable, and a categorical variable indicating rms in above-average enforcement countries. Our second coe cient repeats the same procedure but uses the long-run post-GDPR period instead. Our main coe cients of interest (the triple interactions) measure the short- and long-run di erences in . $_{8C}$for EU rms with above-median strictness relative to those with below-average strictness post-GDPR. Table 5 summarizes the results. The interaction coe cients (although not statistically signi cant) 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 rms in the above-average strictness countries.

4.4 Discussion

Our results so far suggest that EU rms responded to the GDPR by storing less, computing less, and becoming less data-intensive relative to US rms. These results are important for several reasons. First, we provide direct and large-scale evidence that rms comply with the GDPR by signi cantly reducing their data and computation. Second, we show that the GDPR distorts rms' input choices by changing the composition of data and computation used in rm production. Third, the results are not driven by a single industry, by a single country, or exclusively by website rms that are a ected 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 e ect of GDPR is likely to di er across rms because some rms rely on data more heavily than others.

Although these ndings provide insights into the impact of privacy laws on rm behavior and provide direct evidence, they do not o er a comprehensive understanding of rm-speci c economic costs. Such an analysis requires understanding how rms use data in production and the di erent adjustment margins of rms. 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 rms use data and computation in production and how privacy regulations might a ect these decisions. One key consequence of the GDPR is that rms' data costs are a ected. 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:

$$I_{it} = \left[\left(A^{c}_{it} C_{it} \right)^{\frac{\sigma-1}{\sigma}} + D_{it}^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}$$

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 A^{c}_{it} is compute productivity. The parameter $\sigma \in (0, \infty)$ is the elasticity of substitution between data and computing.

Our model includes a firm-specific productivity term, A^{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 A^{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

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

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:

$$\log \frac{\ _{it}}{\ _{it}} = \ + \ \log \frac{p^{d}_{it}}{p^{c}_{it}} + \ \log(A^{c}_{it}) \tag{3}$$

where $\ = \ \log(\ )$. 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 . 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 A^{c}_{it}, 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 inter-

the wedge between the actual cost of data and the total cost that includes complying with GDPR. We model this wedge as rm-speci c because compliance costs will likely be heterogeneous across rms, depending on their size and the types of data they collect. Alternatively, we can also interpret $_{8}$ as each rm's perceived cost of the GDPR, as they may hold di erent beliefs about enforcement or have varying levels of risk aversion that a ect the expected cost of liability in the event of a data breach. We follow the literature and model $_{8}$ as a multiplicative wedge (e.g., Chari et al., 2007; Hsieh and Klenow, 2009).

5.3 Identi cation of Parameters

Our end goal is to estimate two parameters: the wedge introduced by the GDPR ($_{8}$) and the elasticity of substitution between computation and data. To illustrate the potential identi cation problems when estimating $_{8}$ and , consider the rst-order condition in equation (3) after the GDPR for EU rms:

$$\log \frac{{}_{8C}}{{}_{8C}} = ¸ \quad \log \frac{?^{3}_{8C}}{?^{2}_{8C}}$$

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 rm is large enough to a ect list prices with their changes in productivity. These changes, however, are still predictive of the prices that rms 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 speci c data centers for each rm 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 rms 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, rms' location choices are highly persistent over time.

More formally, the shift-share design combines list prices with variation in rms' pre-existing data center location choices. We construct instruments I^{3}_{8C} and I^{2}_{8C} for the data storage and computation prices each rm 8 faces at time C The exposure shares for each service in a given period are calculated as the share of rm 8s usage in a given data center relative to the rm's total demand. This di erential exposure gives us the following equation for the instrument:

$$I^{f2-3g}_{8C} = \tilde{O}_{;2} \; B^{f2-3g}_{8;C-12} \, ?^{f2-3g}_{;C} \tag{7}$$

where $B^{f2-3g}_{8;C-12}$ denotes rm 8s usage share for data center location ; as measured 12 months before C, $?^{f2-3g}_{;C}$ is the price index for each service in location ; at time C, and 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 rm are exogenous to the changes in the ratio of list prices of cloud computing in the rm's historical data center choices, controlling for industry-speci c 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 identi cation would be if idiosyncratic compute productivity shocks are strongly correlated over time after accounting for aggregate industry time trends, and this caused rms

We use $I^2_{8C} \bullet I^3_{8C}$ as an instrument for price ratio $?^3_{8C} \bullet ?^2_{8C}$ 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 rm-speci c compute productivity ($\$^2_8$) and production technology parameters before the GDPR. We also estimate Equation (6) for US industries over the entire sample period, as US rms do not experience regulatory distortion either before or after the GDPR. This allows us to recover the industry-speci c compute productivity trends, $)^2_C$ for US industries.

5.3.2 Second Step: Identi cation 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 ($_8$) 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: $?^3_{8C} = {}^11 ¸ \;\; {}_8{}^o ?^3_{8C}$, where $_8$ re ects the cost of the GDPR. Incorporating this into the rm's input demand, we obtain the following equation:

$$\log \frac{}{}\frac{_{8C}}{_{8C}} = \quad ¸ \quad {}_2 \log \frac{?^3_{8C}}{?^2_{8C}} ¸ \quad {}_2 \log{}^1 1 ¸ \quad {}_8{}^o ¸ \quad {}_2 \log{}^1 \$^{2o}_8 ¸ \quad {}_2 \log{}^1)^{o}_C ¸ \quad {}_2 \log{}^1 \; {}_{8C}{}^o \qquad (8)$$

where $_2$ is the post-GDPR elasticity of substitution. Here, unlike the pre-GDPR input demand equation, the additional term $_8$ a ects the ratio of computing to storage. The higher the cost of the GDPR, $_8$, the more likely rms are to substitute away from data toward computation. In order to use this equation for identifying $_8$, we make the following assumptions:

Assumption 1.2222ll842vi15tiow8 28 -rur(a)842n Cost of83(11.9552 Tf 0 -25.0)842a(22)842reasTf bl18

(Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). The typical approach in that literature assumes that rms have the same production technology. This assumption is needed because otherwise the rm-speci c wedges cannot be distinguished from arbitrary rm-level heterogeneity in production technology. We face the same identi cation problem but take a di erent 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 di er 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 speci c to data input. This identi cation 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 identi cation of wedges requires controlling for aggregate changes in compute productivity. Otherwise, the changes in the computation-to-data ratio of EU rms due to GDPR may be attributed to di erential aggregate trends in compute productivity in Europe. Therefore, we use the estimated post-GDPR industry trend from the US rms to control for industry trends in the EU. In particular, the parallel trends we nd 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:

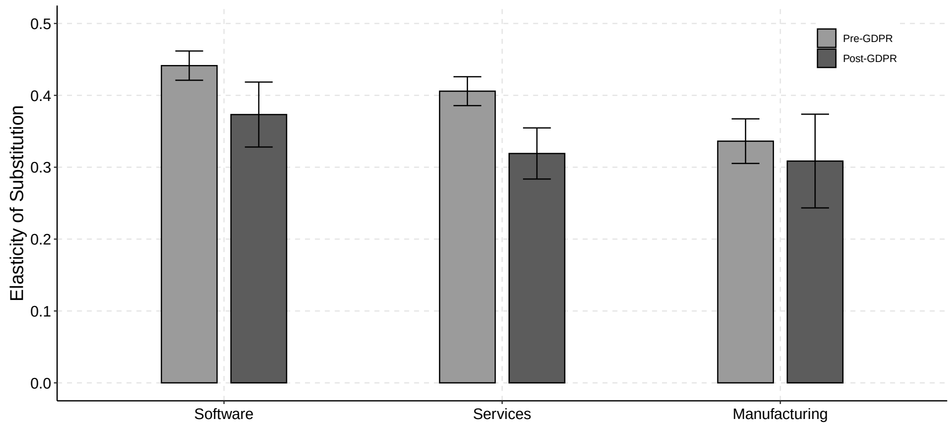
$$\log \frac{8C}{8C} = 2 \,¸\, 2 \log \frac{?^{3}_{8C}}{?^{2}_{8C}} \,¸\, \log^{1}) C^{o} \,¸\, 2 \log^{1}1 \,¸\, 8^{o} \,¸\, \log^{1}\$^{2o}_{8} \,¸\, \log^{1} 8C^{o} \quad (9)$$

where $\$^{2}_{8}$ denotes estimates of compute productivity using pre-GDPR data and $)_{8}$ denotes the estimates of compute productivity trend of the US rms. This equation allows us to estimate our main object of interest ($_{8}$) along with the post-GDPR elasticity of substitution between computing and f 9.18(ticity)y est1EU

Table 6: Elasticity of Substitution Results by Industry

Industry	Software		Services		Manufacturing	
	OLS	IV	OLS	IV	OLS	IV
Elasticity of Substitution ^{1} ^{o}	0.45 (0.02)	0.41 (0.03)	0.45 (0.02)	0.44 (0.04)	0.38 (0.04)	0.34 (0.05)

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



Notes: Figure presents our estimation results of the elasticity of substitution between storage and computing ( ) across industries, and we present separate estimates for the pre- and post-GDPR ($_1$ and $_2$, respectively). Gray bars denote the 95% con dence intervals, and standard errors are calculated using 100 bootstrap repetitions at the rm level.

and associated -statistics. The rst-stage coe cients are positive, indicating a positive relationship between our shift-share instruments and the contemporaneous prices faced by rms. Our results also indicate high -statistics, suggesting that our instruments are strongly correlated with the endogenous variables and that we have a robust rst stage.

The elasticity of coe cient 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 rms 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 rms 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 rms' compute productivity and data-to-compute price ratios is positive; rms 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 rms. 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 ($_8$). 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

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.

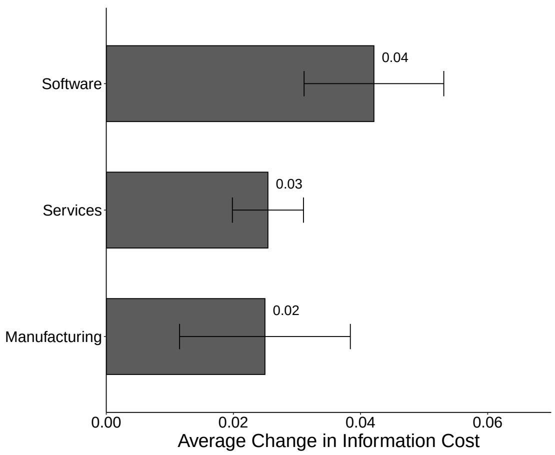
6.3 Cost of Information

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.

.

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 signi cantly 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 considerab42 .Qdar088-eraand 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 ($_{8}$) to the cost of information in the counterfactual with no wedge ($_{8} = 0$). Panel (a) presents the average estimated increase in the cost of information for rms within each industry. Standard errors are calculated using 100 bootstrap repetitions at the rm 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 " rm 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 rm-level information on how rms 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 exible, have common prices, and that rms do not have market power (Foster et al., 2008; Backus, 2020). Using these assumptions, 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 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 ( )	0.04	0.03	0.02
Range of Information Expenditure Share ( )	8.7% - 16.7%	2.9% - 5.0%	2.3% - 3.3%
Resulting Increase in Production Costs ( )	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,15(t)10(or)e thcom5(tiimenu9e0ction.)-2053(at967(com5%)-250(-(e)-277(a)-323

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 Ine ciencies 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 Arti cial Intelligence. Harvard Business Review Press.

Agrawal, A., J. McHale, and A. Oettl (2019). Finding Needles in Haystacks: Arti cial Intelligence and Recombinant Growth. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), The Economics of Arti cial 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 E ect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR. RAND Journal of Economics (Forthcoming).

Arrieta-Ibarra, I., L. Go , D. Jiménez-Hernández, J. Lanier, and E. G. Weyl (2018). Should

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 101(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.

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).

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 (2nd ed.). IT Governance Publishing.

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

Lefrere, V., L. Warberg, C. Cheyre, V. Marotta, and A. Acquisti" (2022). Does Privacy

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. Hu aker, kc cla y, and S. Greenstein (2021). The Impact of the General Data Protection Regulation on Internet Interconnection. Telecommunications Policy 45(2), 102083.

Zolas, N., Z. Kro , E. Brynjolfsson, K. McElheran, D. N. Beede, C. Bu ngton, 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

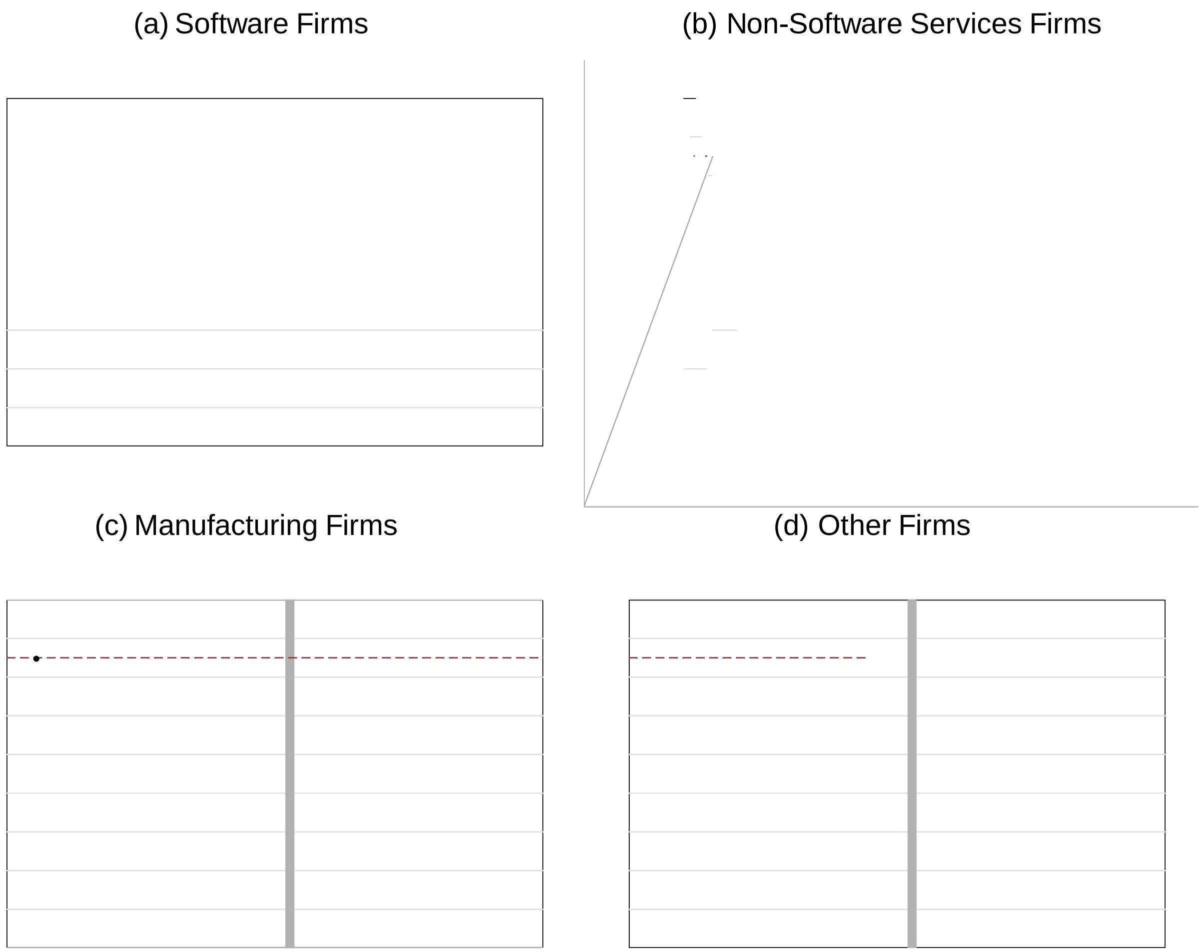
Mert Demirer, Diego Jiménez-Hernández, Dean Li and Sida Peng

Appendix - For Online Publication

Contents

A Additional Exhibits

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



Notes: Figure presents estimates of equation (1) of @ 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 E ect of GDPR on Cloud Inputs
(E ects 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 @ the coe cient on the quarter of the move interacted with our treatment indicator, when the outcome is log computation. The coe cient in the quarter before the GDPR's implementation is normalized to zero. Gray bars represent the 95 percent con dence intervals, and standard errors are clustered at the rm level. Results are broken down by industry.

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 ( ) across industries. We present separate estimates for the pre- and post-GDPR ($_1$ and $_2$, respectively). Standard errors are calculated using 100 bootstrap repetitions.

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 infringements.

B.2 The Compliance Cost of GDPR

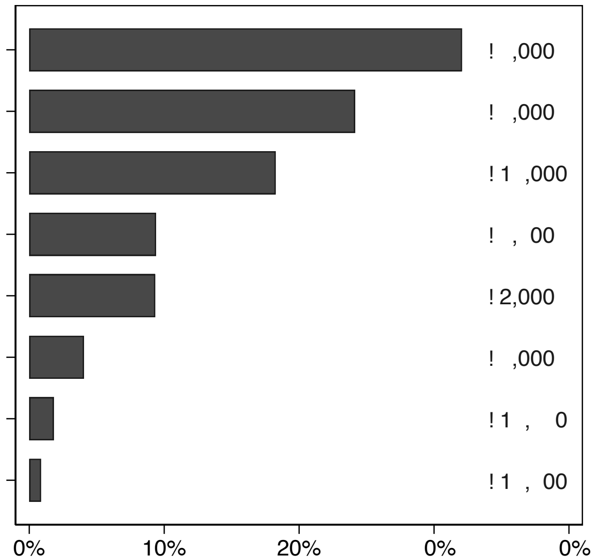
be increasing with the amount of data stored by the rm. Moreover, one can imagine that the probability of a cyberattack could increase with the amount of data. Another related variable cost is cybersecurity insurance. Of the 1,263 organizations surveyed in Ponemon Institute (2019), 31% of respondents purchased insurance covering cyber-risks. Of those insured, 43% had insurance coverage for GDPR nes and penalties.

B.3 Publicly Available GDPR Fine Data

Our primary source of publicly available ne data is a database maintained by CMS Legal Services, a large international law rm that operates in over 40 countries. This data provides an overview of the public nes and penalties that data protection authorities have imposed under the GDPR. Although not all nes are made public, the data on public nes is quite rich, containing the ne amount, the entity being ned, the country of the ne, and the GDPR articles under which the ne was leveled. [53] The database contains more than ¿3 billion in nes levied in the ve years after the implementation of the GDPR. Furthermore, there are primary and secondary sources associated with each of the nes in the database.

For each ne, we scrape the ne amount, the entity that it was levied on, the date, and the reason that the ne was levied. In Figure 1 in the paper, we show the distribution of ne sizes, highlighting that there is considerable variation in the size of the nes. There is also substantial variation in the speci c reasons that nes were levied, and these reasons fall into eight categories: (a) insu cient legal basis for data processing, (b)insu cient involvement of data protection o cer, (c) insu cient technical and organizational mea-

Figure OA-4: Publicly Reported GDPR Fines



Notes: Figure presents the distribution of reasons given for GDPR nes, using the publicly reported ne

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 rms 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 con gured 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 o er hundreds of di erent con gurations, and the user chooses the exact con guration 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

C.1.2 Storage

usage than the same rm's usage in the months immediately preceding and following the month. We also lter these by minimum size change, to ensure that we are not spuriously removing small rms 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 rms whose observations are a ected by the introduction and phaseout of older service models for our provider.

We then construct our sample by conditioning on continuous rm observation for one full year exactly two years before the GDPR. Although the resulting sample of rms is smaller, conditioning on the continuously observed rms does not signi cantly change the share of data that we observe. In fact, these continuously observed rms are responsible for about 90 percent of storage and computation before the GDPR. We present summary statistics on these sets of rms below in Table OA-1. While for con dentiality, we cannot provide direct comparisons between the number of rms 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 rms 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

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 o er 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 nd that the true positive rate is 64 percent, increasing with rm size, and the true negative rate is 66 percent, decreasing with rm 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.

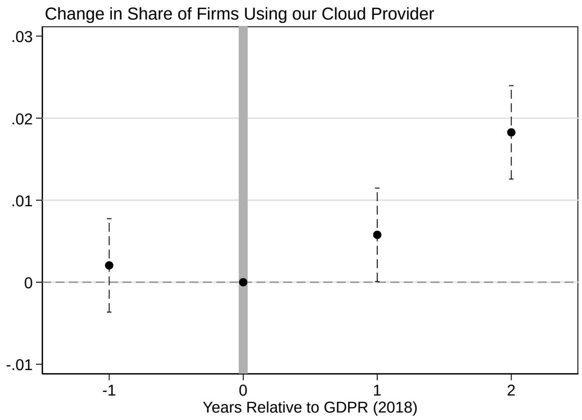
Substitution to Other Cloud Providers Multi-cloud usage where firms get cloud services from multiple cloud computing providers -is common among firms. Industry surveys suggest that 70 percent of cloud users are multi-cloud. Multi-cloud usage could be a potential issue because we observe usage from only 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 nd that the share of rms that are using our cloud provider has moderately increased over time, while the share of rms 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.

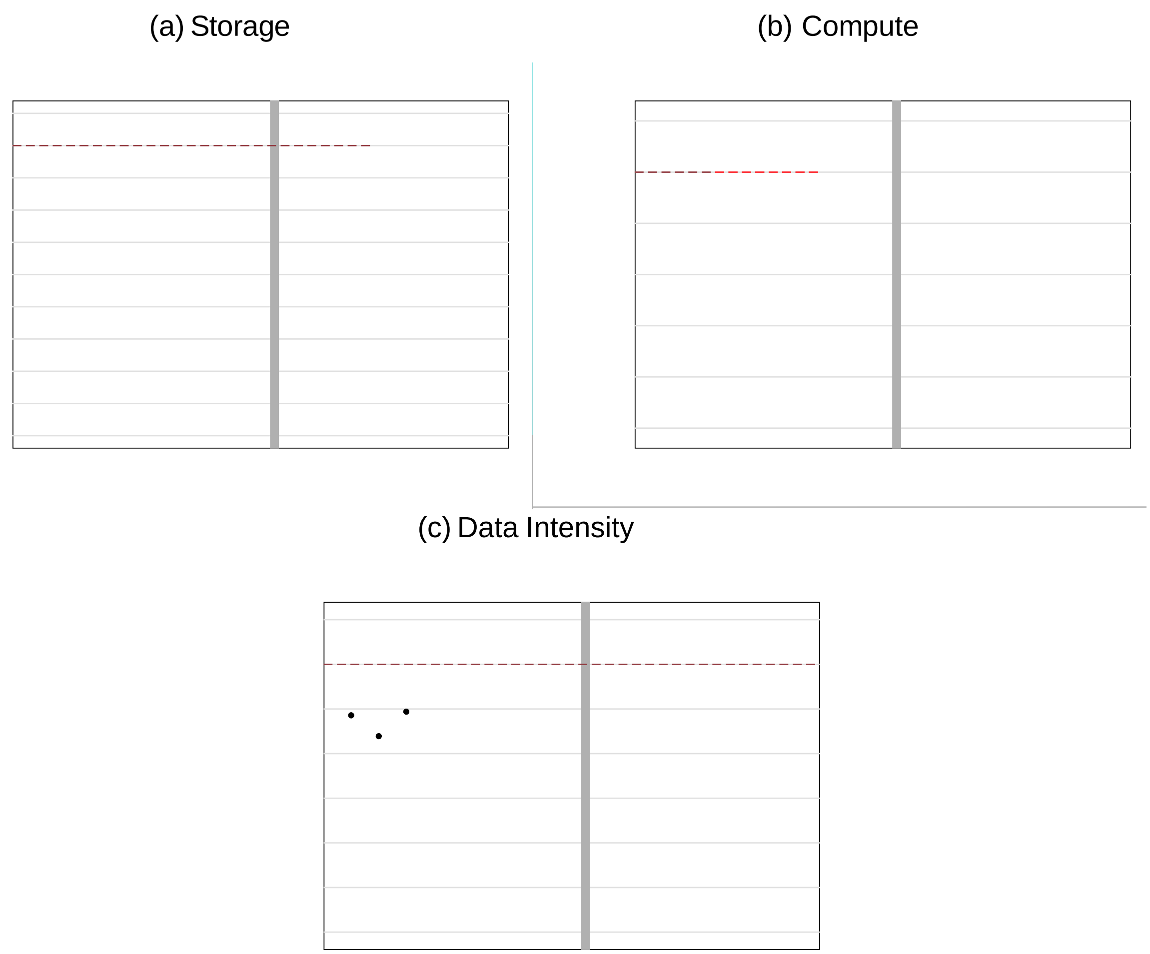
Figure OA-5: Change in Share of Firms Using Cloud Providers in the EU vs the US

(a) Our Cloud Provider



(b)

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



Notes: Figure presents estimates of equation (1) of @ the coe cient on the quarter of the move interacted with our treatment indicator. The coe cient 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 con dence intervals, and standard errors are clustered at the rm level. Sample sizes are presented in Table OA-2. The sample is composed of rms that do not use multiple cloud computing providers.

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

	Storage (1)	Compute (2)	Data Intensity (3)
Short-Run E ect	-0.128	-0.085	-0.061
	(0.020)	(0.019)	(0.023)
Long-Run E ect	-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 ($_1$) and long-run ($_2$) coe cients, which estimate the impact of the GDPR in the rst and second year after the GDPR came into force. Column (1) estimates the e ect on storage. Column (2) estimates the e ect on computation. Column (3) presents estimates of the data intensity. The sample excludes multi-cloud rms as described in Appendix D. Industries are de ned as the ten divisions classi ed by SIC codes, with the addition of a "software" division, which we carve out of the services division and de ne through SIC codes 7370 - 7377. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we de ne size decile as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the rm 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.

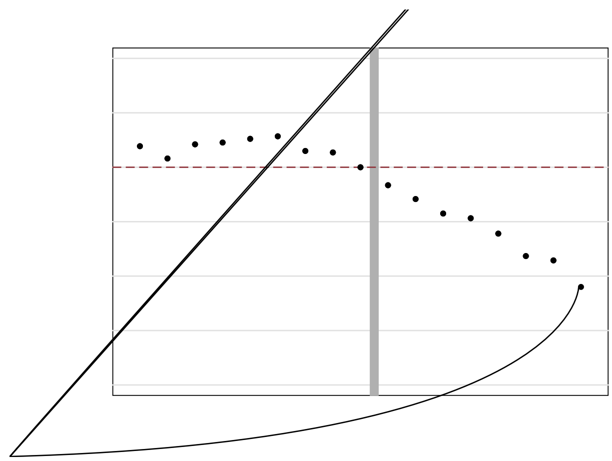
Substitution to Traditional IT Next, we consider that rms might use both traditional IT and cloud computing. To the extent that we cannot observe traditional IT usage, declines in cloud computing may re ect re-allocations towards traditional IT rather than true declines in computing. While increasing cloud computing adoption rates suggest that this margin may not play an important role, we consider the possibility that post-GDPR, European rms might have changed allocation between two ITs di erently from the US rms.

This would invalidate our identi cation arguments for the e ects of compute and storage, though it should not necessarily a ect 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 rms 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 rms. In Appendix Table OA-4 and Figure OA-8, we actually nd larger e ects for these rms rather than smaller e ects. This suggests that the observed declines in computing and storage are unlikely to be driven by substitution

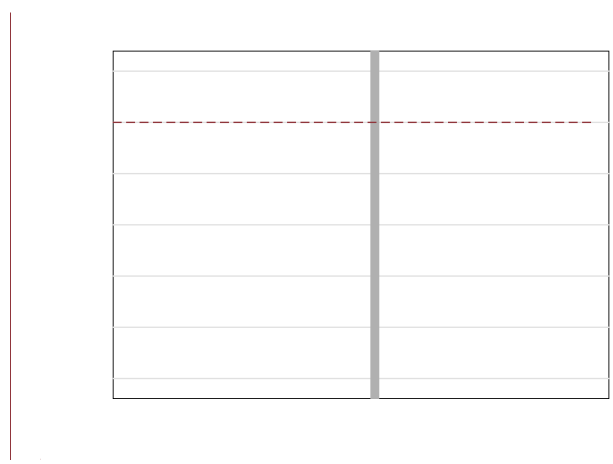
Figure OA-7: Event Study Estimates of the E ect of GDPR on Cloud Inputs
(Balanced Panel Estimates)

(a) Storage

(b) Compute



(c) Data Intensity



Notes: Figure presents estimates of equation (1) of @ the coe cient on the quarter of the move interacted with our treatment indicator. The coe cient 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 con dence intervals, and standard errors are clustered at the rm level. Sample sizes are presented in Table OA-2. The sample is a balanced panel, and details can be found in Appendix Section D.

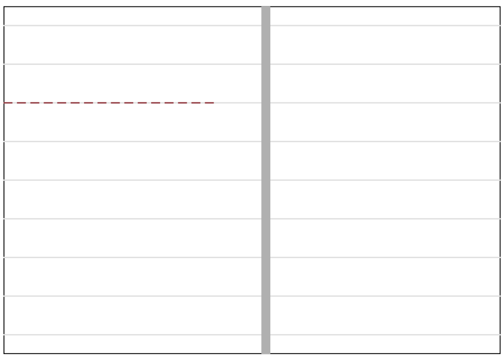
Table OA-3: Short- and Long-Run E ects of GDPR
(Balanced Panel Estimates)

	Storage (1)	Compute (2)	Data Intensity (3)
Short-Run E ect	-0.221	-0.115	-0.046
	(0.024)	(0.020)	(0.027)
Long-Run E ect	-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

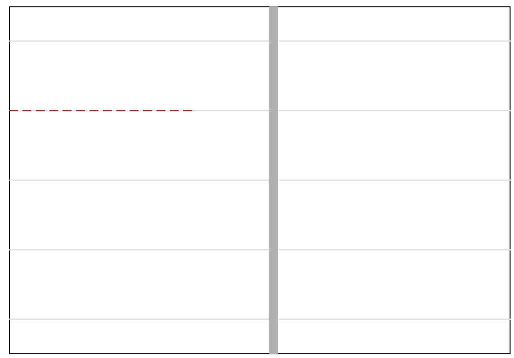
Notes: Table presents estimates of equation (2) of the short-run ($_1$) and long-run ($_2$) coe cients, which estimate the impact of the GDPR in the rst and second year after the GDPR came into force. Column (1) estimates the e ect on storage. Column (2) Data Intensity

Figure OA-8: Event Study Estimates of the E ect of GDPR on Cloud Inputs (Start-Up Firms)

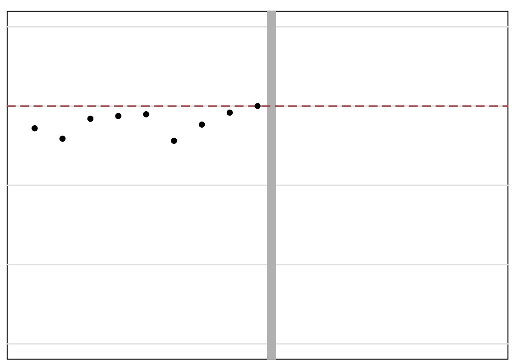
(a) Storage



(b) Compute



(c) Data Intensity



Notes: Figure presents estimates of equation (1) of @ the coe cient on the quarter of the move interacted with our treatment indicator. The coe cient 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 con dence intervals, and standard errors are clustered at the rm level. Sample sizes are presented in Table OA-4. The sample is composed of start-up rms, where start-ups are labeled according to a de nition internal to the cloud provider.

D.2 Price Changes

One natural channel through which the GDPR may have a ected rms 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 rms. 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 di erential 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 speci cation using paid prices as the outcome. We nd no evidence of signi cant di erential price changes.

Figure OA-9: Event Study Estimates of the E ect of GDPR on Cloud Inputs
(E ects on Paid Prices)

(a) Storage Prices

(b) Compute Prices

Notes: Figure presents estimates of equation (1 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

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 ${}_1$ and ${}_2$, 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 ${}_1$ and ${}_2$, 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 (${}_1$) and long-run (${}_2$) 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 De nitions We also discuss the robustness of our analyses in Section 4 to alternative sample de nitions. In particular, we show that our estimated coe cients are relatively stable when estimated when conditioning on a di erent window of pre-GPDR usage, and when using a larger and more inclusive de nition of rms" 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 rms 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 E ects of GDPR
(Alternative Pre-GDPR Usage Windows)

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

Notes: Table presents estimates of equation (2) of the short-run ($_1$) and long-run ($_2$) coe cients, which estimate the impact of the GDPR in the rst and second year after the GDPR came into force. Column (1) shows our baseline speci cation. Column (2) conditions on observing rms for the year before GDPR (instead of two years before). Column (3) restricts the sample to rms continuously observed for the full two years before GDPR. Industries are de ned as the ten divisions classi ed by SIC codes, with the addition of a "software" division, which we carve out of the services division and de ne through SIC codes 7370 - 7377. Pre-GDPR size deciles are measured thirteen months before the GDPR. For data intensity, we de ne size decile as the interaction between storage and compute terciles when measured in the period. Standard errors are clustered at the rm level.

Finally, we consider using a larger and more inclusive de nition of rms". Per Appendix C, we de ne rms in our baseline sample by requiring that there be either internal or external information on the rm's industry and country. In this larger sample, we drop the restriction that we must observe the rm's industry. Because there is no industry information, we amend the speci cation in equation (2) so that xed e ects do not vary by industry. Appendix Table OA-9 below presents our estimates using this alternative

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 @ 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 rms produce according to the following production function:

$$H_{8C} = 5^{1}\text{-}\ {}_{8C}\ {}_{8C}\$_{8C}$$

where ${}_{8C}$ represents information, -_{8C} is a vector of other observed inputs, and $\$_{8C}$ represents unobserved inputs. We assume that the information is produced according to the following technology:

$${}_{8C} = \$^{2}_{8C}{}^{1}\ {}_{8C}\ {}_{\text{¸}}\ {}_{8C}{}^{1\bullet}\ \bullet$$

Without loss of generality, we can normalize $= 1$ due to the homotheticity of the CES production function: $\$^{2}_{8C}{}^{1}\ {}_{8C}\ {}_{\text{¸}}\ {}_{8C}{}^{1\bullet} = \$^{2}_{8C}{}^{\bullet}\ {}^{1}\ {}_{8C}\ {}_{\text{¸}}\ {}_{8C}{}^{1\bullet}$.

We assume that rms choose variable inputs to minimize the cost of production taking prices as given, a necessary condition for pro t maximization. We also assume that rms take productivity $\$^{2}_{8C}$ as given which follows an exogenous process. This cost minimization problem can be written as:

$$\min_{{}_{8C}\ {}_{8C}}\ ?^{2}_{8C}\ {}_{8C},\ ?^{3}_{8C}\ {}_{8C},\ ?^{G}_{8C}\text{-}^{E}_{8C} \quad \text{s.t.} \quad 5^{1}\text{-}\ {}_{8C}\ {}_{8C}\$_{8C} \quad .^{¢}_{8C}$$

where $.^{¢}_{8C}$ is the target level of production and -^{E}_{8C} denotes variable inputs. The FOCs with respect to ${}_{8C}$ and ${}_{8C}$ can be written as:

$${}_{8C}5^{1}\text{-}\ {}_{8C}\ {}_{8C}\$_{8C}\ \$^{2}_{8C}{}^{1}\ {}_{8C}\ {}_{\text{¸}}\ {}_{8C}{}^{1\bullet 1}\ {}^{1o}\ {}_{8C}{}^{1}\ {}^{1o}F^{2}_{8C} = ?^{2}_{8C}$$

$${}_{8C}5^{1}\text{-}\ {}_{8C}\ {}_{8C}\$_{8C}\ \$^{2}_{8C}{}^{1}\ {}_{8C}\ {}_{\text{¸}}\ {}_{8C}{}^{1\bullet 1}\ {}^{1o}\ {}_{8C}{}^{1}\ {}^{1o} = ?^{3}_{8C}$$

where ${}_{8C}$ is the Lagrange multiplier. Taking the ratio of the two FOCs, we obtain:

$$\frac{{}_{8C}}{{}_{8C}}{}^{1}\ {}^{1o}\$^{2}_{8C} = \frac{?^{2}_{8C}}{?^{3}_{8C}}$$

Taking the logarithm and rearranging the terms yields:

$$ {}^{1}1 \quad {}^{o}\log \frac{8C}{8C} \quad \log{}^{1}\$ \, {}^{2}_{8C}{}^{o} = \log \frac{?^{3}_{8C}}{?^{2}_{8C}} \tag{13}$$

By using = 1•¹ 1 º, we can obtain Equation (3) as presented in the main text

$$ \log \frac{8C}{8C} = \log \frac{?^{3}_{8C}}{?^{2}_{8C}} \, ¸ \quad \log{}^{1}\$ \, {}^{2}_{8C}{}^{o}• \tag{14}$$

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 rms might require software engineers to process data. To illustrate this scenario, we consider a nested CES form where data and computation are nested:

$$ {}_{8C} = \$ \, {}^{2\,1}_{8C} \; 8C \; ¸ \quad {}_{8C}{}^{E•} \; ¸ \; ! \, ! \, {}^{E}_{8C} \, {}^{1•E} $$

Taking the rst-order conditions with respect to $_{8C}$ and $_{8C}$ we obtain:

$$ {}_{8C}2 \, 5^{1}\text{-} \; 8C \; 8C \$ \, 8C \; \$ \, {}^{2\,1}_{8C} \; 8C \; ¸ \; {}_{8C}{}^{E•} \; ¸ \; ! \, ! \, {}^{E}_{8C}{}^{1•E \; 1} \; \$ \, {}^{2\,1}_{8C} \; 8C \; ¸ \; {}_{8C}{}^{E•1} \; 1^{o} \; {}^{1}_{8C} \, 1^{o} F^{2}_{8C} = ?^{2}_{8C} $$

$$ {}_{8C}2 \, 5^{1}\text{-} \; 8C \; 8C \$ \, 8C \; \$ \, {}^{2\,1}_{8C} \; 8C \; ¸ \; {}_{8C}{}^{E•} \; ¸ \; ! \, ! \, {}^{E}_{8C}{}^{1•E \; 1} \; \$ \, {}^{2\,1}_{8C} \; 8C \; ¸ \; {}_{8C}{}^{E•1} \; 1^{o} \; {}^{1}_{8C} \, 1^{o} = ?^{3}_{8C} $$

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

$$ \frac{8C}{8C}{}^{1} \; 1^{o} \, \$ \, {}^{2}_{8C} = \frac{?^{2}_{8C}}{?^{3}_{8C}} • $$

Therefore, the information production function can accommodate labor. It is important to note that this result relies on the speci c 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 $?_2$ and $?_3$ to denote the price of computation and data, respectively. We also use $ in place of $ 2. From the rst-order conditions, we obtain:

$$1 = \frac{?_2}{?_3}\frac{1}{\$} \; 1 \; - \tag{15}$$

which yields:

$$?_3^{\bullet 1} \; 1^{o} \; \$ \; ^{\bullet 1} \; 1^{o} = ?_2^{\bullet 1} \; 1^{o} \; \bullet$$

Adding $?_2^{\bullet 1} \; 1^{o}\$$ to both sides of Equation (15) and simplifying yields:

$$? \;_2 \; ?_2^{\bullet 1} \; 1^{o}\$ \; ¸ \; \$ \; ^{\bullet 1} \; 1^{o}?_3^{\bullet 1} \; 1^{o} \; 1^{\bullet} = ?_2^{\bullet 1} \; 1^{o} \; ¸ \; \$ \; 1^{\bullet} \; \bullet \tag{16}$$

Similarly, adding $\$^{1\bullet 1} \; 1^{o}?_3^{\bullet 1} \; 1^{o}$ to Equation (15) and simplifying yields:

$$? \;_3 \; ?_2^{\bullet 1} \; 1^{o}\$ \; ¸ \; \$ \; ^{\bullet 1} \; 1^{o}?_3^{\bullet 1} \; 1^{o}o1^{\bullet} = \$^{1\bullet 1} \; 1^{o}?_3^{\bullet 1} \; 1^{o} \; ¸ \; \$ \; 1^{\bullet} \; \bullet \tag{17}$$

Adding Equations (16) and (17) and using $= \; ¸ \; \$ \; 1^{\bullet}$, we arrive at:

$$? \;_3 ¸ \; ? \;_2 \; \$^{1\bullet} = \$^{1\bullet 1} \; 1^{o}?_3^{\bullet 1} \; 1^{o} \; ¸ \; ?_2^{\bullet 1} \; 1^{o} \; ^{1} \; 1^{o\bullet} \; \bullet$$

To derive the cost of information, we need to express the sum $? \;_3 ¸ \; ? \;_2$ as a function of and prices. We do this by isolating the sum on one side of the equation:

$$\begin{aligned} ? \;_3 ¸ \; ? \;_2 &= ?_3^{\bullet 1} \; 1^{o} \; ¸ \; \$^{1\bullet 1} \; ?_2^{\bullet 1} \; 1^{o} \; ^{1} \; 1^{o\bullet} \\ &= {}^{1}\$^{o} \; \frac{1}{?_2} \; ^{1} \; ¸ \; \frac{1}{?_3} \; ^{1} \; !^{1\bullet 1} \; 1^{o} \; \bullet \end{aligned}$$

Finally, using $= 1^{\bullet 1}1 \; ^{o}$, we arrive at the desired cost function equation.

$$^{1} \; _{8C} - ?^{8}_{8C} = {}_{8C}{}^{1}\$^{2\,o}_{8C} \; \frac{1}{?^2_{8C}} \; ^{1} \; ¸ \; \frac{1}{?^3_{8C}} \; ^{1} \; !^{1\bullet 1} \; 1^{o} \; \bullet$$

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, , as:

$$^{1} - ? - ^{o} = ?_2 \quad ^{1} - ? - ^{o} \,\text{¸}\, ?_3 \quad ^{1} - ? - ^{o}$$

where $^{1} - ? - ^{o}$ and $^{1} - ? - ^{o}$ 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 is obtaining by di erentiating both sides with respect to :

$$\frac{d}{d} = ?_2 \frac{dC}{d} \,\text{¸}\, ?_3 \quad \text{¸}\, ?_3{}^{1}1 \,\text{¸} \quad {}^{o}\frac{dD}{d}$$

Multiplying both sides by • we obtain:

$$\frac{d}{d}\,\overline{} = ?_2 \frac{dC}{d}\,\overline{} \,\text{¸} \quad \frac{?_3}{} \,\text{¸}\, ?_3{}^{1}1 \,\text{¸} \quad {}^{o}\frac{dD}{d}\,\overline{}$$

Rearranging terms, and multiplying the rst term by • , and the third by • we get

$$\frac{d}{d}\,\overline{} = \frac{?_3}{} \,\text{¸}\, \frac{?_2}{} \; \frac{dC}{d}\,\overline{} \,\text{¸}\, \frac{?_3{}^{1}1 \,\text{¸} \quad {}^{o}}{} \; \frac{dD}{d}\,\overline{}$$

and nally recognizing that the terms in parenthesis are the expenditure shares B_3 and B_2, and the terms in squared parenthesis are the elasticities, we get to Equation (

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

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 rm has exposure to di erent locations and pays di erent prices in each location due to variations in list prices and rm-speci c discounts. We denote rm speci c price indices by $?^{3}_{8C}$ and $?^{2}_{8C}$ for data and computation, respectively. This price could be endogenous because the rm may negotiate lower prices or change its exposure to di erent locations based on productivity. To instrument for these prices, we use the list prices of storage in location ;, given by $?_{;C}$. This price is plausibly exogenous to changes in rm productivity because, after controlling for industry-speci c trends, no rm is likely to a ect list prices in a speci c 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 $I^{3}_{8C} = \sum_{;} B^{3}_{8C\,12;} ?^{3}_{;C}$ for storage and I^{2}_{8C} for computation calculated similarly. Finally, we use $I^{2}_{8C} \bullet I^{3}_{8C}$ to instrument for $?^{2}_{8C} \bullet ?^{3}_{8C}$ in the production function estimation. Since we need the 12 months lagged exposure of each rm, we lose the rst 12 months of observations when implementing this instrumental variable strategy.

F.4 Estimation Details

Our identi cation strategy relies on the assumptions that the industry-speci c cloud productivity trend in Europe would have followed that of US rms in the absence of GDPR, and that rm-speci c compute technology does not change post-GDPR. To operationalize these assumptions, we follow a two-step estimation strategy

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

$$\log \frac{8C}{8C} = \quad ¸ \quad {}^{*(}_{1} \log \frac{?^{3}_{8C}}{?^{2}_{8C}} ¸ \quad {}^{*(}_{1} \log{}^{1}\$^{2o}_{8} ¸ \quad {}^{*(}_{1} \log{}^{1}){}^{2o}_{C} ¸ \quad {}^{*(}_{1} \log{}^{1} \; {}_{8C} \qquad (18)$$

When estimating this equation, we normalize to zero because it is not separately identi- ed from the mean of $\$^{2}_{8}$. We also normalize $)^{2}_{1}$ to 1 so that productivity trend is relative to the initial period. Since, by assumption, the US rms have not been exposed to GDPR, this equation identi es the industry-speci c compute productivity trends, or $)^{2}_{C}$ in Equation (9). By Assumption (2), the EU industries follow the same trend and we use the estimated $)^{2}_{C}$ for EU rms.[58] Next, we estimate the same equation using EU rms only with pre-GDPR data. This estimation identi es $\$^{2}_{8}$ in Equation (9) because there is no distortion

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

before GDPR to estimate ${}^*_1$. We report the associated elasticity estimates in Figure 4 as the pre-GDPR elasticity of substitution estimates.

These rst-step estimations identify provide us with $\$^2_8$ and $)_G$ Using those we nally estimate Equation (9):

$$\log \frac{8C}{8C} = {}_2 ¸ \; {}^*_2 \; \log \frac{?^3_{8C}}{?^2_{8C}} ¸ \log^1) {}^o_C ¸ \; {}^*_2 \; \log^1 1 ¸ \; {}_8^o ¸ \log^1 \$^{2o}_8 ¸ \log^1 \; {}_{8C}^o •$$

by constructing the right-hand side variable. We report ${}^*_2$ as the post-GDPR elasticity of substitution estimates in Figure 4. To estimate the wedge, ${}_8$, we subtract $\log^1 \$^{2o}_8$ from the estimated xed e ects in Equation (9) (after accounting for ${}^*_2$). We report the estimates of ${}_8$ in Figure 5. To account for uncertainty in rst-step estimates in standard errors, we follow a bootstrap procedure with 100 repetitions. We resample rms 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 $\$^2_8$, we calculate the cost of information by setting ${}_8$ 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- rm level.

To do the decomposition presented in Equation 11, we calculate the cost share of data every period using rm's data input demand and prices. The direct e ect is obtained by multiplying the data share with rm-speci c wedges. The second term (rm re-adjustment) is obtained by subtracting the direct e ect 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- rm level.

F.5 Identi cation Intuition for the Firm-Speci c Wedges

Having outlined our estimation strategy in the previous subsection, we now explain how

ourHaving(es.)1es.es.es.es.-44J/F11disobtainF11disdemand an pby13.746w1(in)-280(rs)15(es3eng we subtract

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.

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:

.¢ ?– = ?! ! .¢ ?– ¸ ? .¢ ?– ¸ ? .¢ ?– •

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:

. = min ! ——— •

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

! = .¢

= .¢

= .¢

In this case, the cost function is therefore linear in prices, and a percentage increase in the cost of information causes an B8C percentage increase in the cost of production.

Linear Production Function

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

. = ! ¸ ¸

In the interior case where rms were previously producing with non-zero capital or non-zero labor, cost minimization immediately implies that a 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 e ects of a percentage increase in the cost of information for a Cobb-Douglas production function given by

. = !

First-order conditions imply the following information demand function:

= .¢ 1 ¸ ¸ ? ¸ ¸ ? ¸ ¸ ?! ¸ ¸

This immediately implies that a percentage increase in? induces a = h ¹1 ¸ º ¸ ¸ ¸ 1 i percentage decrease in .[61] Next, we note that rst-order conditions imply that a share of total rm costs will be spent on information:

= ? .¢ ?− / .¢ ?− •

Using the change in information expenditure resulting from the increase in information prices and the decrease in derived above, we have that a percentage increase in? will lead to a percentage increase in production costs, where = ¹1 ¸ º 1.[62]

61 For marginal changes, using log transformations and taking derivatives yields %log

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 to know how an increase in the cost of information translates to production costs. We note that 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:

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 speci cally Flexera's