Data, Privacy Laws & Firm Production: Evidence from GDPR

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

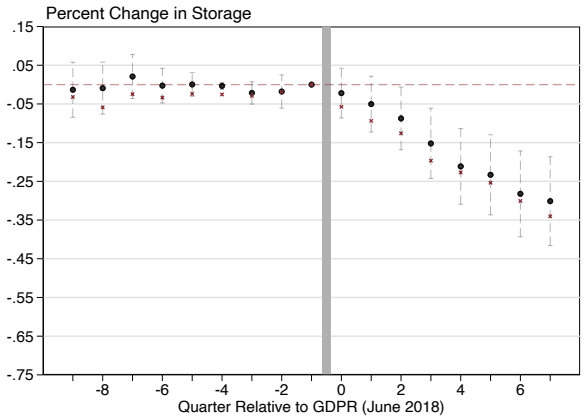
Appendix - For Online Publication

Contents

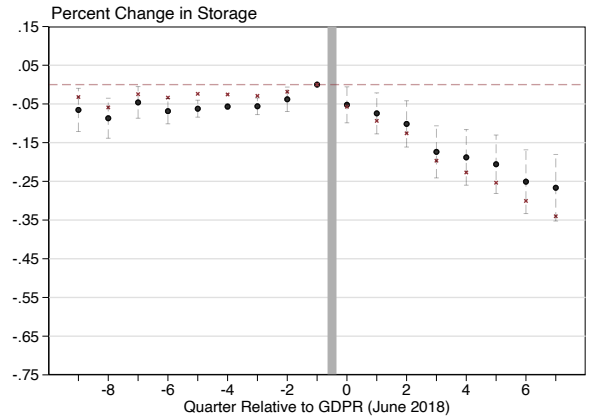
A Additional Exhibits

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

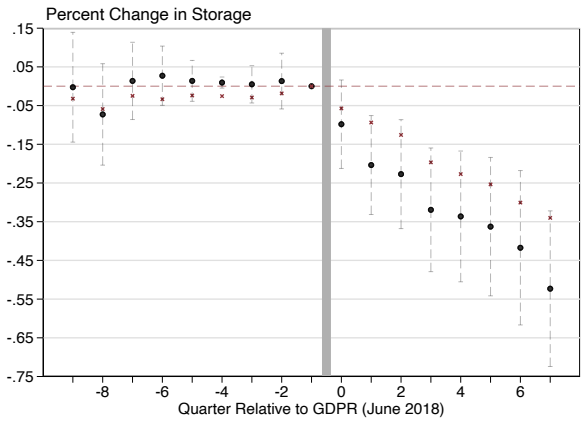
(a) Software Firms



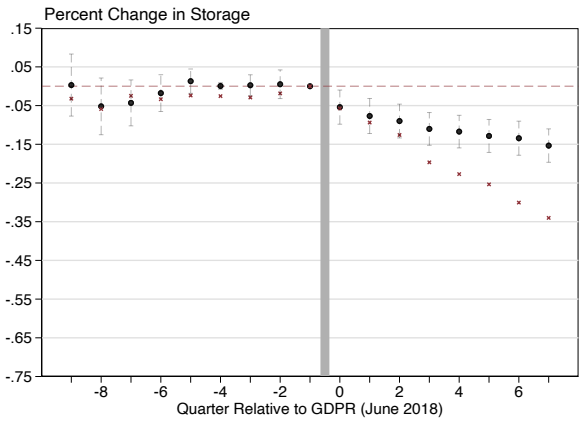
(b) Non-Software Services Firms



(c) Manufacturing Firms



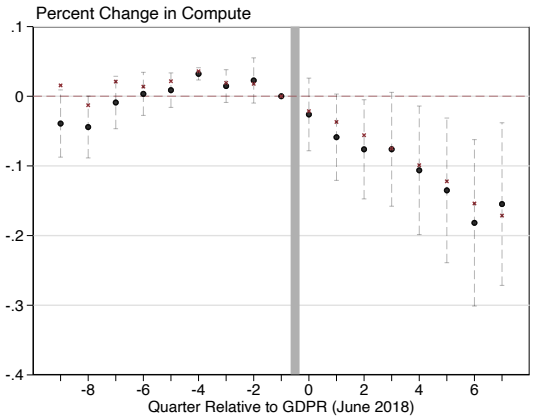
(d) Other Firms



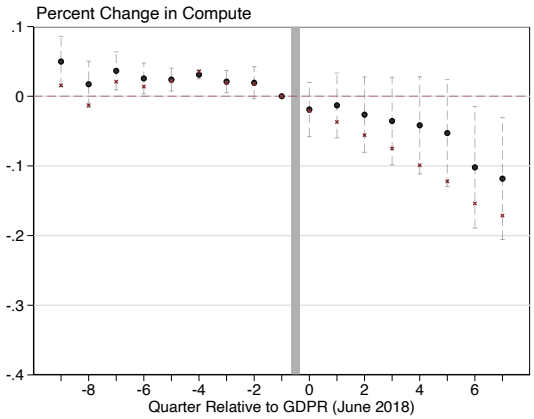
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator, when the outcome is log storage. The coefficient in the quarter before the GDPR's implementation is normalized to zero. Dashed bars represent the 95 percent confidence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

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

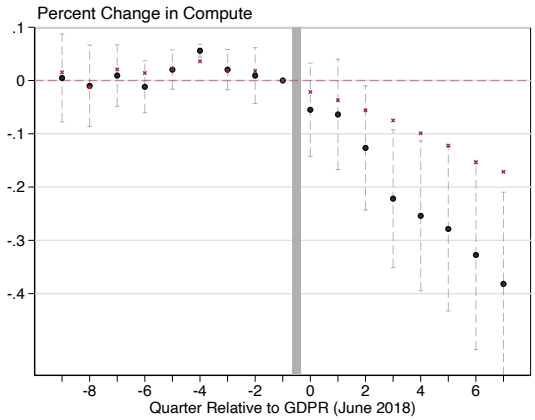
(a) Software Firms



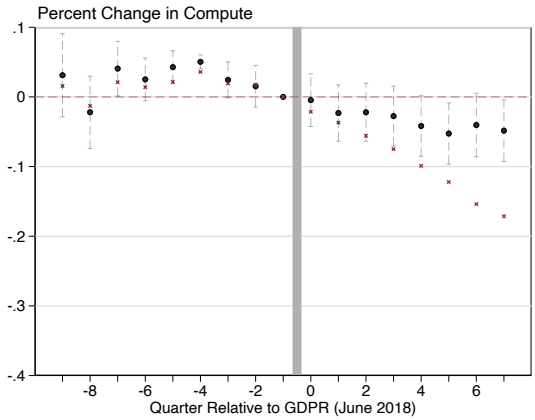
(b) Non-Software Services Firms



(c) Manufacturing Firms

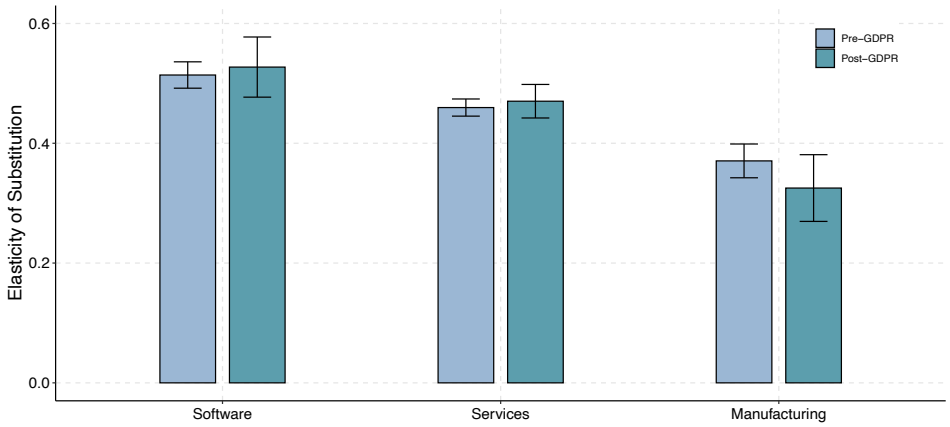


(d) Other Firms



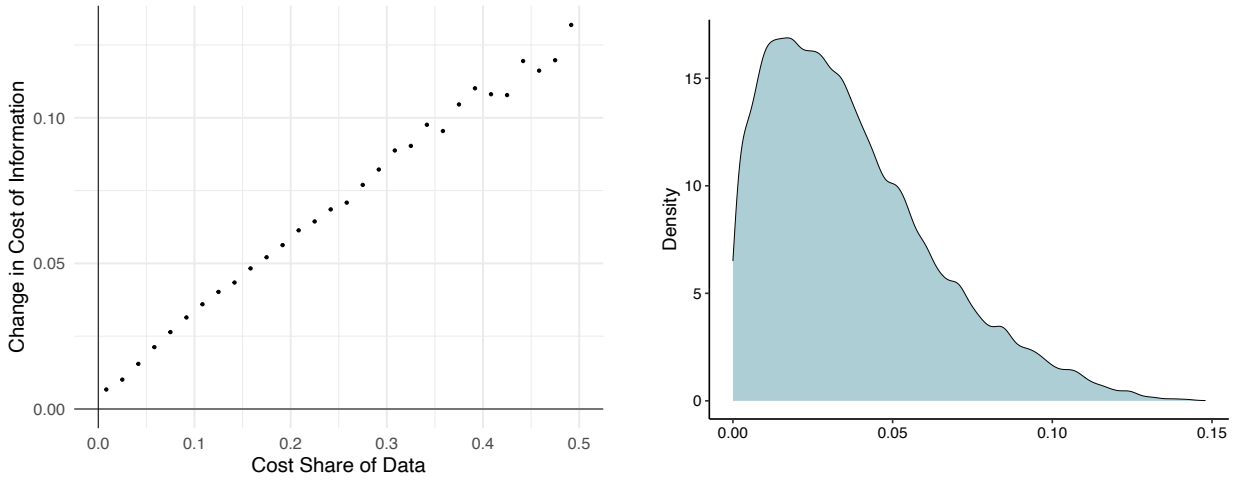
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator, when the outcome is log computation. The coefficient in the quarter before the GDPR's implementation is normalized to zero. Dashed bars represent the 95 percent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

Figure OA-3: Elasticity of Substitution Between Storage and Computing for US Firms



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

Figure OA-4: Additional Results on Information Cost



(a) Avg. Change in Info. Cost by Data Share **(b)** Firm Re-Adjustment Margin

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

B The Impact of the GDPR on Firms

B.1 GDPR Summary

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

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

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

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

Preparing a Data Protection Impact Assessment (Article 35). If an intended processing activity, especially one involving new technologies, is likely to result in a high risk to the rights and freedoms of data subjects, then firms must conduct a Data Protection Impact Assessment (PIA) to identify and implement appropriate measures to mitigate privacy

risks. The PIA should be conducted at the start of a project so that all stakeholders are aware of any potential privacy risks. The PIA should include the following components: (i) a systematic description of the purposes and planned processing operations, including the controller's legitimate interests (if applicable); (ii) an assessment of the necessity and proportionality of the processing in relation to the purpose; (iii) an assessment of the risks to the rights and freedoms of the data subjects; and (iv) the measures planned to address these risks.

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

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

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

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

infringements.

B.2 The Compliance Cost of the GDPR

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

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

B.2.1 Fixed and Sunk Costs

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

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

firms have appointed multiple DPO (Hughes and Saverice-Rohan, 2017), indicating that DPO could be a variable cost for large firms.

B.2.2 Variable Costs

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

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

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

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

Liabilities The maximum penalties under the GDPR include fines of up to €20 million or 4% of the company's global annual revenue, whichever is greater. However, the actual penalty amount is determined by the nature and severity of the violation and is likely to

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

B.3 Publicly Available GDPR Fine Data

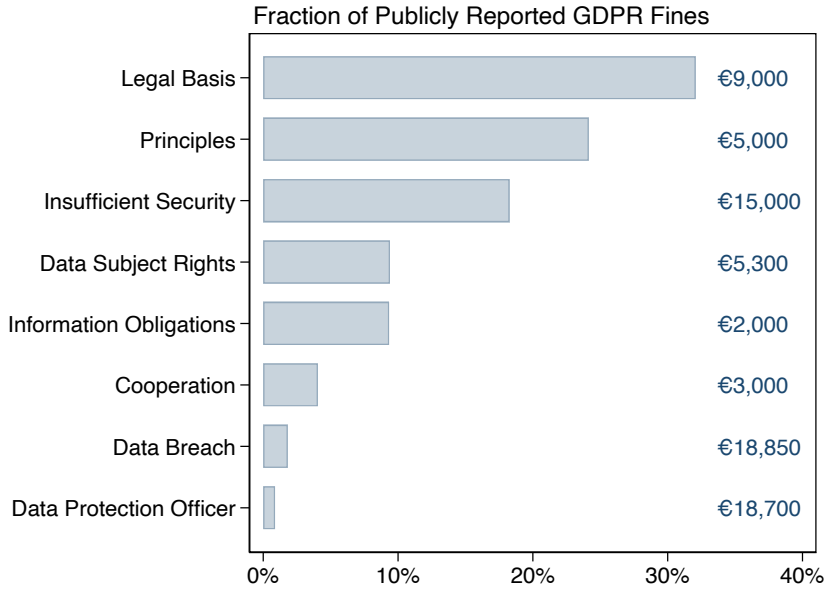
Our primary source of publicly available fine data is a database maintained by CMS Legal Services, a large international law firm that operates in over 40 countries. This data provides an overview of the public fines and penalties that data protection authorities have imposed under the GDPR. Although not all fines are made public, the data on public fines is quite rich, containing the fine amount, the entity being fined, the country of the fine, and the GDPR articles under which the fine was leveled.[50] The database contains more than €3 billion in fines levied in the five years after the implementation of the GDPR. Furthermore, there are primary and secondary sources associated with each of the fines in the database.

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

In Figure OA-5, we show the share of fines that were levied under each reason and the median fine size conditional on the reason. Perhaps unsurprisingly, data security concerns result in the largest types of fines. The median fines for insufficient information security and insufficient notification of data breaches are €15,000 and €18,850 respectively, while the median fines for non-cooperation and insufficient fulfillment of information obligations

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

Figure OA-5: Publicly Reported GDPR Fines



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

are €3,000 and €2,000 respectively. Overall, the distribution of the reasons given for the publicly available GDPR fines suggests that fines may be levied against firms for a variety of reasons.

C Data Appendix

C.1 Cloud Computing Details

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

C.1.1 Computation

Firms that require computation on the cloud typically opt for virtual machines (VMs). VMs are a type of cloud computing "compute" product that allows users to create and manage virtual machines instead of maintaining their own physical hardware.[51] These VMs run on virtualized infrastructure provided by a cloud computing provider and can access software and computing resources. These machines are typically fully customizable and controlled by the user. Cloud computing VMs can be configured in various ways. Some of the features of virtual machines that can be customized include memory, storage, networking options, CPU, operating system, and the location of the data center that hosts the VM. Cloud computing providers offer hundreds of different configurations, and the user chooses the exact configuration when requesting a VM.

In our paper, we use the number of CPU cores in a virtual machine as the key measure of computation outcome because this is the key vertical VM characteristics that determines computing performance. We note, however, that this approach does not take into account heterogeneity in other characteristics, such as how much memory and network capability is combined with the number of cores.

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

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

C.1.2 Storage

Cloud providers offer a wide range of storage products that can be used for various purposes, including storing and managing data, backing up and recovering data, and archiving data for long-term retention. These products can be categorized into two types: disk storage and database storage. Disk storage provides physical hardware where firms can store a wide variety of data, including operating system files, applications, documents, and multimedia files. Disk storage can include different physical configurations, such as Hard Disk Drives (HDDs) and Solid-State Drives (SSDs), as well as Storage Area Networks. Disk storage can also differ based on other characteristics such as upload and download speed. Databases, on the other hand, are collections of structured data that are hosted and managed in a cloud computing environment by a cloud provider. The differentiation of databases refers to the various types of databases available and their specific features and characteristics, such as MySQL, NoSQL, Oracle, and PostgreSQL.

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

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

C.2 Sample Selection and Cleaning

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

Next, we clean the data by removing outlying observations. In order to tag a firm as an outlier, we require that we observe the firm's usage in the months immediately preceding and following a given month. We define outliers as large and sudden temporary spikes or temporary dips. These are months where a firm's usage is either twenty times more or less

usage than the same firm's usage in the months immediately preceding and following the month. We also filter these by minimum size change, to ensure that we are not spuriously removing small firms with more volatile usage. This cleaning removes less than 0.1 percent of observations. We also worked with internal employees to conduct some minor cleaning to remove a small fraction of firms whose observations are affected by the introduction and phaseout of older service models for our provider.

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

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

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

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

C.3 Aberdeen Sample

Aberdeen is a market research firm that provides valuable information on firms' hardware and software investments. They gather this information from various sources. Every year, they survey a sample of senior IT executives about their software and hardware usage and extrapolate this information to non-surveyed firms. Additionally, they conduct large-scale

data collection efforts, such as web scraping job postings and purchasing customer lists from vendors to identify software choices. Our understanding is that technology adoption information comes only from the latter source. This data also includes sales, the number of employees, industry, and a DUNS number, and these firm characteristics are sourced from Duns & Bradstreet. Our sample of Aberdeen data covers the period from 2015 to 2021 at the annual level. The data from Aberdeen has been used to study digitization and technology adoption (Graetz and Michaels, 2018; Tuzel and Zhang, 2021).

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

C.3.1 Match Procedure Between Aberdeen and Cloud Data

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

We use the Jaro-Winkler (JW) distance to match names, which considers the number of transpositions and the number of matching characters between two strings. Intuitively, strings with more characters in common and requiring fewer transpositions for one string to be contained within the other have lower distances. For the same number of character matches and transpositions, the JW distance is smaller for strings that match the first characters of the strings.[52]

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

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

Aberdeen dataset (either by using the parents or the subsidiaries' name). We sequentially match using the following criteria and say that two firms are a match if both:

1. Share the same DUNS number, or
2. Share the same website, or
3. Are in the same postal code and the name distance is less than 0.1, or
4. Are in the same city and the name distance is less than 0.08, or
5. Are in the same state and the name distance is less than 0.07, or
6. Are in the same country and the name distance is less than 0.065, or
7. Are in the same region (e.g., EU) and the name distance is less than 0.045.

Suppose a firm in the cloud computing data has multiple matches in the Aberdeen data. In that case, we hierarchize based on the same order as we list our criteria above.[53] 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

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

position to offer a partial answer to this question because we possess internal data from one of the largest cloud providers and cross-check Aberdeen data for this provider.

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

D Robustness Checks

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

D.1 Substitution to Other Providers

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

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

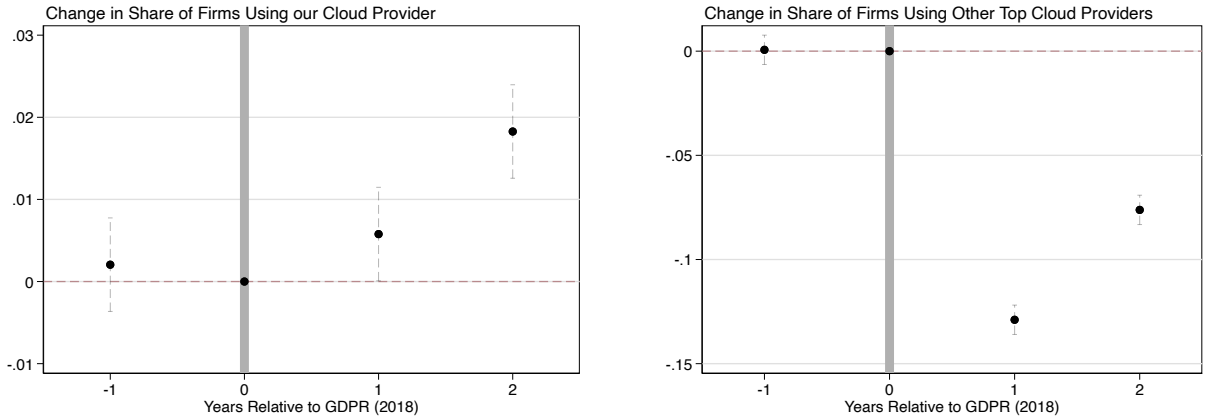
In particular, we might attribute a decline in cloud storage and computing to firms simply switching their cloud usage to other providers. We note, however, that firms that conduct both storage and computing are likely to do both with the same provider because data cannot be stored with one provider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

We attempt to address the identification challenge to our storage and computing results with three additional exercises. First, we bring an external dataset, Aberdeen, that provides information on firms' technology adoption and which vendors they get cloud services from. Using this dataset, we look at our provider's share of firms that receive services from each of the top cloud providers in Europe and US before and after GDPR and plot

them in Appendix Figure OA-6. We find that the share of firms that are using our cloud provider has moderately increased over time, while the share of firms using the other cloud providers has decreased. Thus, we do not expect the relative attractiveness of the cloud provider that we observe to have decreased after GDPR.

Figure OA-6: Change in Share of Firms Using Cloud Providers in the EU vs the US

(a) Our Cloud Provider **(b)** Other Cloud Providers



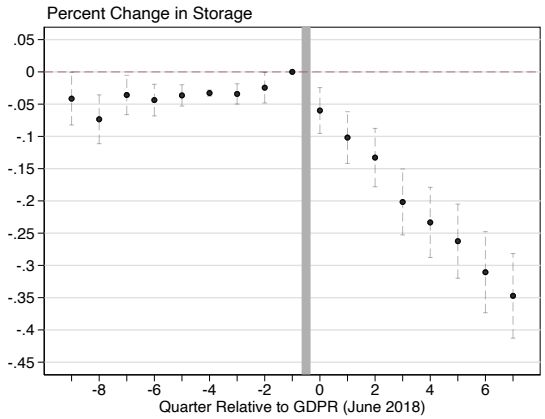
Notes: Figure presents estimates of the difference in the share of firms who use different cloud providers in the EU vs the US. The data source is Aberdeen (formerly known as Harte Hanks). The dependent variable on the left panel is equal to one if a firm uses the cloud provider that we study in this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

Second, we identify single cloud firms using Aberdeen again and estimate our empirical specification using only these firms. In Appendix C.3.2 we assess the reliability of Aberdeen data to identify these single-cloud firms and show that Aberdeen data provides useful information to detect single-cloud firms. Appendix Table OA-2 and Appendix Figure OA-7 present our estimates using this sample, which are quite similar to our baseline estimates across all outcomes. As discussed in the paper's main body, it is unlikely that the declines we observe are simply driven by substitution in usage to other providers.

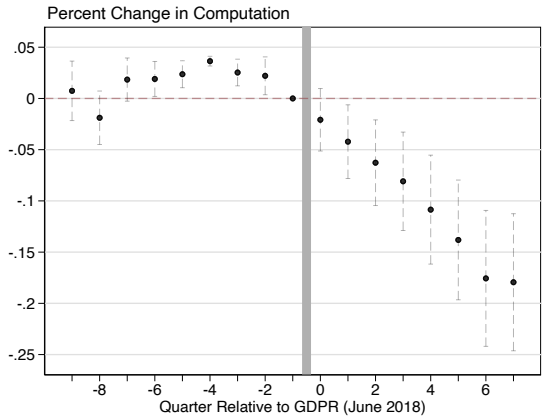
Finally, as discussed in Appendix B.1, the GDPR is likely to make multi-cloud usage more difficult. Thus, switching between cloud providers is more likely to occur on the *extensive* margin rather than the *intensive* margin. Thus, any cloud usage declines in a sample of firms that continuously use our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

Figure OA-7: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Excluding Multi-Cloud Firms)

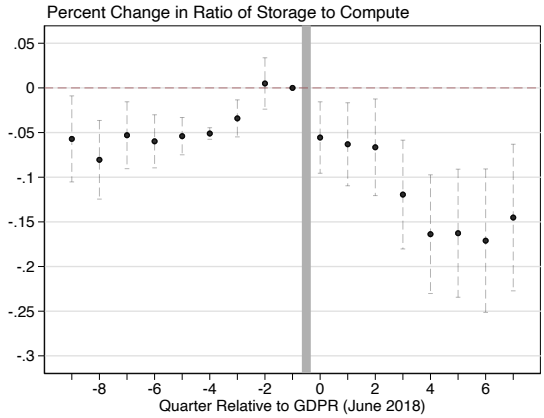
(a) Storage



(b) Compute



(c) Data Intensity



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

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

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

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

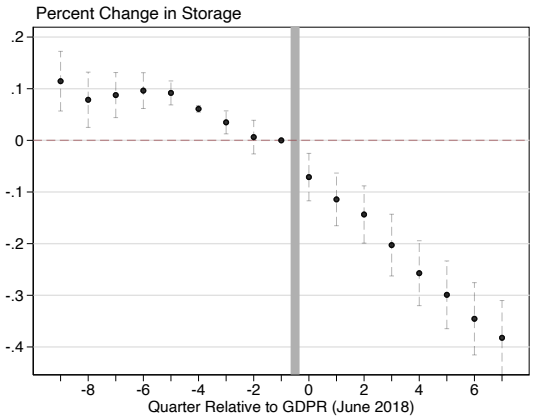
of a large decrease in both compute and storage alongside a decrease in data intensity. Thus, the results from our balanced panel in Appendix Table OA-3 and Appendix Figure OA-8 suggest that the declines in computation and storage we observe are not driven by switching between providers.

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

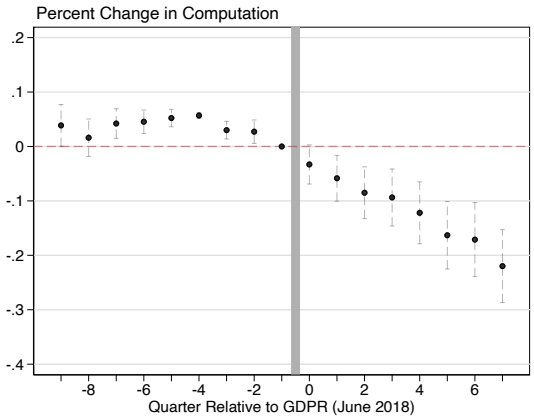
This would invalidate our identification arguments for the effects of compute and storage, though it should not necessarily affect the results on data intensity. To provide a robustness check for this, we focus on start-ups, which are unlikely to be switching to traditional IT. These are young software firms for which the upfront costs of traditional IT make it unlikely for them to switch towards these technologies as they are likely to face larger costs than e.g., more established firms. In Appendix Table OA-4 and Figure OA-9, we actually find larger effects for these firms rather than smaller effects. This suggests that the observed declines in computing and storage are unlikely to be driven by substitution

Figure OA-8: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Balanced Panel Estimates)

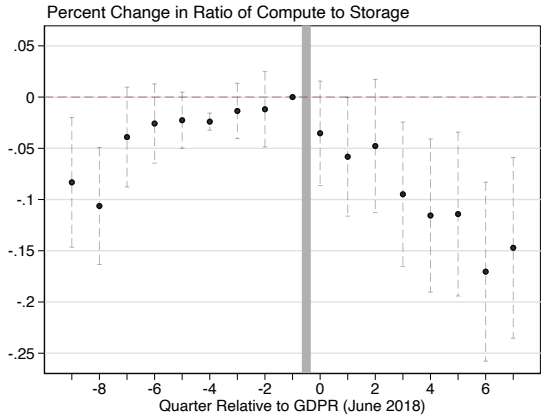
(a) Storage



(b) Compute



(c) Data Intensity



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

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

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

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

to traditional IT.

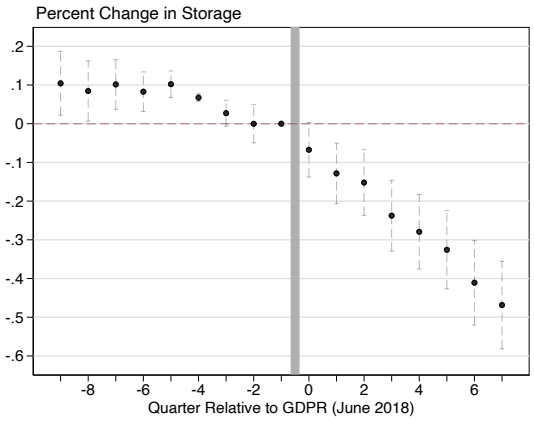
Table OA-4: Short- and Long-Run Effects of GDPR
(Start-Ups Firms)

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

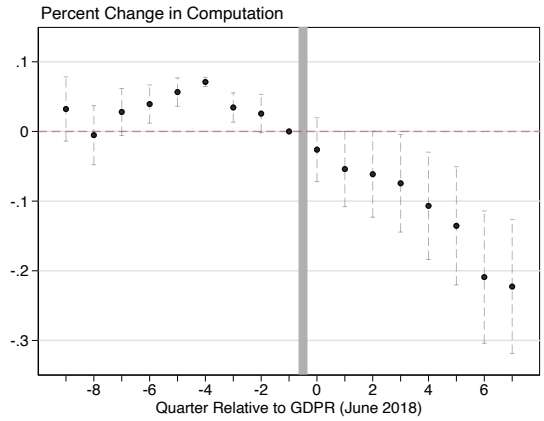
Notes: Table presents estimates of equation (2) of the short-run (δ_1) and long-run (δ_2) coefficients, which estimate the impact of the GDPR in the first and second year after the GDPR came into force. Column (1) estimates the effect on storage. Column (2) estimates the effect on computation. Column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

Figure OA-9: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Start-Up Firms)

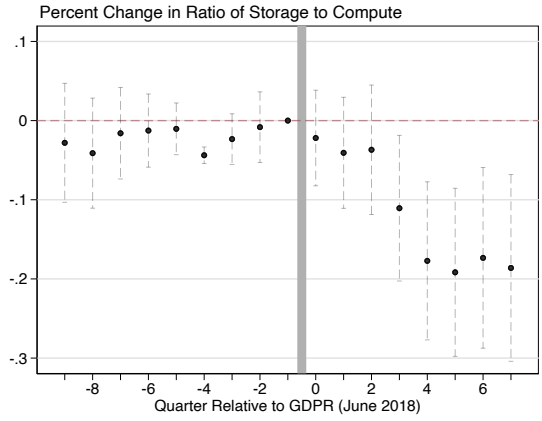
(a) Storage



(b) Compute



(c) Data Intensity



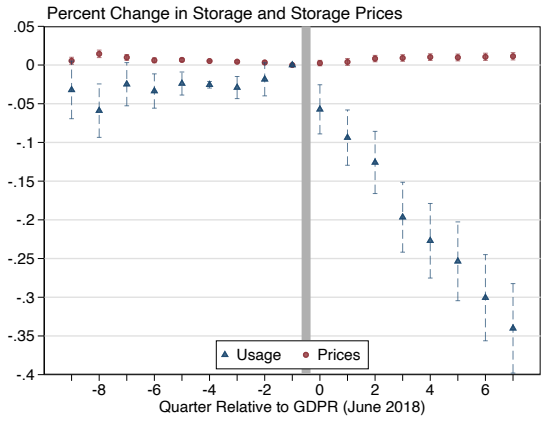
Notes: Figure presents estimates of equation (1) of β_q, the coefficient on the quarter of the move interacted with our treatment indicator. The coefficient in the quarter before the GDPR's implementation is normalized to zero. The outcome in each subpanel is denoted by the subpanel title. Dashed bars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

D.2 Price Changes

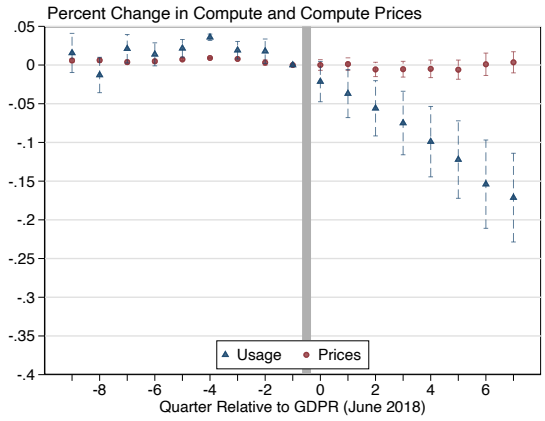
One natural channel through which the GDPR may have affected firms is through price changes in cloud computing. This would suggest our results might capture pricing responses by cloud providers rather than the GDPR's direct impact on firms. For example, if cloud computing providers increase their prices in the European Union relative to the United States, this could confound our estimates. While conversations with internal employees suggest that there were no explicit pricing responses to the passage of the GDPR, we also examine the data for evidence of any differential pricing trends between the EU and the US, either in listed or paid prices. Appendix Figure OA-10 presents our results when we estimate our event study specification using paid prices as the outcome. We find no evidence of significant differential price changes.

Figure OA-10: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Effects on Paid Prices)

(a) Storage Prices



(b) Compute Prices



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

D.3 Websites and Cookie Collection

One of the most salient aspects of the GDPR is the requirement that firms receive consent for the collection of data. This is particularly important in the case of websites and cookies: post-GDPR, websites that need to collect personal information must get explicit consent. As studied by Aridor et al. (2022), there may also be selection in terms of which consumers

choose to opt out of data collection and how valuable the remaining data is.

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

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

D.4 Additional Robustness Exercises

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

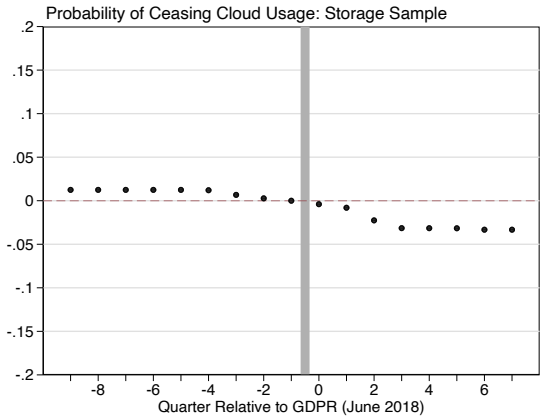
	Storage (1)	Compute (2)	Data Intensity (3)
Short-Run Effect	-0.073	-0.059	-0.063
	(0.013)	(0.013)	(0.015)
Long-Run Effect	-0.151	-0.113	-0.117
	(0.018)	(0.020)	(0.022)
Observations	2,224,810	1,097,922	756,996
US Firms	34,876	18,037	10,807
EU Firms	31,622	15,004	10,299

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

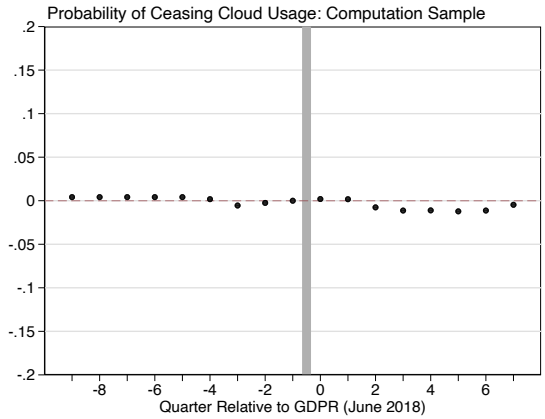
Extensive Margin Although Appendix Table OA-3 suggests that our baseline estimates are similar when we use a balanced panel of firms, we also directly examine whether the GDPR caused differential attrition between firms in the European Union and the United States. We study this using the following same specification but replacing the outcome variable with an indicator for whether the firm has exited our sample. We present these results in Appendix Figure OA-11.

Figure OA-11: Event Study Estimates of the Effect of GDPR on Cloud Inputs (Differential Attrition)

(a) Storage Sample



(b) Compute Sample



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

E Technical Appendix

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

E.1 First-Order Conditions

Assume that firms produce according to the following production function:

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

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

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

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

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

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

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

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

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

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

Taking the logarithm and rearranging the terms yields:

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

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

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

E.2 Including Labor in Information Production Function

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

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

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

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

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

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

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

E.3 Derivation for Cost of Information

In this section, we derive the formula for the cost of information given by Equation (10). To ease notation, we drop the subscript and use p_c and p_d to denote the price of computation

and data, respectively. We also use ω in place of ω^c. From the first-order conditions, we obtain:

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

which yields:

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

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

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

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

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

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

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

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

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

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

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

E.4 Cost of Information Decomposition

In this section, we derive the formula for the decomposition of the cost of information. We drop all subscripts to ease notation and start by substituting the values for the cost

minimizing information cost, CI^*, as:

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

where $C^*(I, p, \lambda)$ and $D^*(I, p, \lambda)$ are the arguments of the cost-minimizing function. We will remove the function arguments to ease out notation even more. The total derivative with respect to λ is obtaining by differentiating both sides with respect to λ:

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

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

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

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

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

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

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

F Model Estimation Details

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

F.1 Cloud Computing Pricing

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

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

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

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

F.2 Price Index Construction

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

To obtain firm-specific price indices, we simply calculate the unit price paid by the firm by dividing the monthly total spending on compute and storage by the total quantity of

compute and storage, respectively. This gives us firm-specific compute and storage price indices, which can vary either because of the discounts negotiated by firms or variation in location-specific prices. We divide the price of storage by the price of computation to obtain a firm-specific storage-to-compute price ratio. Since this ratio involves some outliers due to small values in the dominator, we winsorize these variables by the top and bottom 2 percentiles. We also construct the storage-to-compute ratio for each firm and apply the same winsorization procedure.

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

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

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

F.3 Instrumental Variable Strategy

Our instrumental variable strategy relies on the assumption that firms' choice of data center location is persistent. This assumption is based on the fact that the cost of moving large datasets from one data center to another is typically high. The cost of moving data to another data center in cloud computing can depend on several factors, including the amount of data being transferred, the distance between the source and destination data centers, and the pricing policies of the cloud service provider (García-Dorado and Rao, 2015). Some cloud service providers may charge a fee for data transfer, and there may be additional costs associated with data migration, such as network bandwidth charges, storage costs, and downtime or disruption to services during the migration process.[54] Even though the specific costs and risks of data migration will depend on the migration plan

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

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

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

F.4 Estimation Details

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

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

$$\log\left(\frac{C_{it}}{D_{it}}\right) = \gamma + \sigma_1^{US}\log\left(\frac{p^d_{it}}{p^c_{it}}\right) + \sigma_1^{US}\log(\omega^c_i) + \sigma_1^{US}\log(\phi^c_t) + \sigma_1^{US}\log(\eta_{it}), \tag{18}$$

When estimating this equation, we normalize γ to zero because it is not separately identified from the mean of ω^c_i. We also normalize ϕ^c_1 to 1 so that productivity trend is relative to the initial period. Since, by assumption, the US firms have not been exposed to GDPR, this equation identifies the industry-specific compute productivity trends, or $\hat{\phi}^c_t$ in Equation (9). By Assumption (2), the EU industries follow the same trend and we use the estimated $\hat{\phi}^c_t$ for EU firms.[55] Next, we estimate the same equation using EU firms only with pre-GDPR data. This estimation identifies $\hat{\omega}^c_i$ in Equation (9) because there is no distortion

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

before GDPR to estimate σ_1^{EU}. We report the associated elasticity estimates in Figure 4 as the pre-GDPR elasticity of substitution estimates.

These first-step estimations identify provide us with $\hat{\omega}_i^c$ and $\hat{\phi}_t$. Using those we finally estimate Equation (9):

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

by constructing the right-hand side variable. We report σ_2^{EU} as the post-GDPR elasticity of substitution estimates in Figure 4. To estimate the wedge, λ_i, we subtract $\log(\hat{\omega}_i^c)$ from the estimated fixed effects in Equation (9) (after accounting for σ_2^{EU}). We report the estimates of λ_i in Figure 5. To account for uncertainty in first-step estimates in standard errors, we follow a bootstrap procedure with 100 repetitions. We resample firms with replacement in each industry-continent group and apply the entire estimation procedure.

We use Equation (10) to estimate the change in the cost of information, with results reported in Section 6.3. For the estimated ω_i^c, we calculate the cost of information by setting λ_i to its estimated value and 0, which gives us the change in the cost of information due to GDPR. Since prices change over time, we calculate this change in information cost at every observed price point and report the distribution at the month-firm level.

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

F.5 Identification Intuition for the Firm-Specific Wedges

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

Consider two firms in the same industry, one in the EU (k) and one in the US (j) with the same levels of firm-level compute productivity $\omega_k^c = \omega_j^c$. For simplicity (to not carry

terms around), assume both firms have the same time-varying shocks (i.e., $\log \eta_{kt} = \log \eta_{jt}$ for all t).[56] Subtracting the pre-GDPR first order condition (Equation 6) of the US firm from the EU firm equation in a period $\underline{t}$ before GDPR implies that:

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

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

Analogously, focusing on a period $\bar{t}$ after GDPR was enacted, we can use the post-GDPR identifying equation (Equation 8) in a similar fashion as before (focusing on the same two firms) to obtain:

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

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

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

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

Equation (21) provides useful intuition about what λ_i, the post-GDPR wedge, is intended to capture. Loosely speaking, the wedge captures the variation in the shift in the compute intensity (across EU and US firms, before and after GDPR) that is not explained by changes in the shift in the relative prices, or by pre- and post-GDPR differences in the elasticity of substitution between compute and storage across comparable EU and US firms.[57] Given the above equation, one would intuitively expect firms that face larger

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

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

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

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

Interestingly, Equations (21) and (20) also show that level changes in C_{it} and D_{it} are not enough to identify λ_i. Note that we cannot infer that firms with larger responses in *levels* would have larger (or smaller) wedges. In fact, to rationalize the level responses of compute and storage, one would need additional assumptions over the full production function. To explain the responses in levels, we would need to construct a model that incorporates the elasticity of substitution between information and other traditional inputs (e.g., capital and labor), which we intentionally refrain from doing in this paper.

G Effects on Production Costs

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

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

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

where θ is the percentage increase in the information cost.

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

Leontief Production Function

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

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

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

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

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

Linear Production Function

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

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

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

Cobb-Douglas Production Function

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

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

First-order conditions imply the following information demand function:

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

This immediately implies that a θ percentage increase in p^I induces a $\delta = \left[(1+\theta)^{-\frac{\alpha+\beta}{\gamma+\alpha+\beta}} - 1\right]$ percentage decrease in I^*.[58] Next, we note that first-order conditions imply that a γ share of total firm costs will be spent on information:

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

Using the change in information expenditure resulting from the θ increase in information prices and the δ decrease in I^* derived above, we have that a θ percentage increase in p^I will lead to a ζ percentage increase in production costs, where $\zeta = (1+\theta)^{\gamma} - 1$.[59]

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

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

CES Production Function

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

$$Y_i = \nu_i \left[\alpha I_i^{\frac{\bar{\rho}-1}{\bar{\rho}}} + (1-\alpha) M_i^{\frac{\bar{\rho}-1}{\bar{\rho}}} \right]^{\frac{\bar{\rho}}{\bar{\rho}-1}}.$$

where ν_i represents firm-specific productivity, a_i represents firm-specific information intensity in production, and $\bar{\rho}$ denotes the elasticity of substitution between information and non-information inputs. We drop the firm-specific subscripts moving forward for notational simplicity.

Next, we note that because $M(L, K, \cdot)$ exhibits constant returns to scale, the linear prices of labor and capital – p_L and p_K – imply a linear unit cost for the intermediate non-information aggregate M. We denote that unit cost by p_M.[60] This therefore yields the unit cost function

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

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

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

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

[60] Deriving the explicit formula for the unit cost of the intermediate good yields $p_M = \frac{1}{\gamma} \left(\beta^{\sigma} p_L^{(1-\sigma)} + (1-\beta)^{\sigma} p_K^{(1-\sigma)} \right)^{\frac{1}{1-\sigma}}$.

$$\begin{aligned}\frac{c\left((1+\theta)p_I, p_M\right)}{c\left(p_I, p_M\right)} &= \left(\frac{(1+\theta)^{1-\bar{\rho}}\alpha^{\bar{\rho}}p_I^{1-\bar{\rho}} + (1-\alpha)^{\bar{\rho}}p_M^{1-\bar{\rho}}}{\alpha^{\bar{\rho}}p_I^{1-\bar{\rho}} + (1-\alpha)^{\bar{\rho}}p_M^{1-\bar{\rho}}}\right)^{\frac{1}{1-\bar{\rho}}} \\ &= \left(\frac{(1+\theta)^{1-\bar{\rho}}\left(\frac{\alpha}{1-\alpha}\right)^{\bar{\rho}}\left(\frac{p_I}{p_M}\right)^{1-\bar{\rho}} + 1}{\left(\frac{\alpha}{1-\alpha}\right)^{\bar{\rho}}\left(\frac{p_I}{p_M}\right)^{1-\bar{\rho}} + 1}\right)^{\frac{1}{1-\bar{\rho}}} \\ &= \left((1+\theta)^{1-\bar{\rho}}s_I^* + 1 - s_I^*\right)^{\frac{1}{1-\bar{\rho}}}.\end{aligned}$$

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

G.2 Estimating Key Calibration Parameters

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

First, we combine various data sources to suggest a reasonable range for the information cost share. We provide these estimates in Table OA-10. Next, we discuss each of those data sources separately. Finally, we discuss mapping estimates from Lashkari et al. (2023) of the elasticity of substitution between IT and non-IT inputs into our setting.

Aberdeen

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

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

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

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

Industry Surveys

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

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

We use point estimates of the elasticity of substitution between IT and non-IT inputs from Lashkari et al. (2023) to proxy for the elasticity of substitution between information and non-information inputs. We focus on the micro-elasticities provided in the text rather than the macro-elasticities, which reflect general equilibrium forces and reallocation between firms. We use their industry-level elasticities from a non-homothetic CES specification. Estimates for the manufacturing industry are provided directly. We map the "information

and communication technology" industry to software, and we construct an estimate for the elasticity in the non-software services industries by taking a weighted average of the relevant industries for which estimates were provided in the online appendix.

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

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

We use data from the US census to compute the software industry share of the service sector as we could not find any reliable estimates for the EU. To compute this, we use the 2019 SUSB Annual Data Tables by Establishment Industry provided by the US Census Office. We compute the software industry share by dividing employment in the software industry by the total employment in the service industry and get 7.53%.[61] We use this number to proxy the EU size of the software industry.

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

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