

Value of Data to U.S. Public Firms: Evidence from GDPR*

Po-Yu Liu[†]

March 11, 2024

Abstract

This paper empirically quantifies the value of data to U.S. public firms by using the General Data Protection Regulation (GDPR), an EU privacy law, as a natural experiment that exogenously shocks data processing activities. Analyzing a sample of 2,371 firms over seven years, I find that treated firms — those with larger legal departments — reduce data processing activities by 11% relative to control firms post-GDPR, and experience a 1.2% drop in sales. If the exogenous reduction in data processing contributes fully to the sales decline, it implies a 1% increase in data processing can raise sales by 0.11%, quantifying the value of data. Additional evidence such as treated firms' non-increasing SG&A and IV regressions supports the same interpretation that data processing is the most likely channel between GDPR and sales. Heterogeneity analysis reveals that among treated firms, GDPR's impacts on data and sales coincide in firms with lower share of software engineers or workers located in EU. The coinciding impacts of GDPR on data and sales further support the positive value of personal data.

*I thank Alan Kwan and Tse-Chun Lin for their valuable feedback and suggestions. I also thank seminar participants at The University of Hong Kong, National Taiwan University, and National Chengchi University for many helpful comments and discussions.

[†]HKU Business School, The University of Hong Kong. Email: liupoyu@connect.hku.hk.

1 Introduction

Data are essential to modern business processes and forecasting models. Financial media even call data the new oil of our era. Firms use data to anticipate customer demands, optimize inventories, and deliver personalized advertisements ([Veldkamp, 2023](#)). Understanding the economic return of data can help us optimally invest in data infrastructure, efficiently price data-related products, or determine whether a data company's economic influence triggers anti-trust concerns. However, measuring the economic value of data is often difficult. Data lack standard economic units, and their value can vary for firms with different goals. Moreover, most corporate data are byproducts of business activities, making these data carry zero accounting costs.

This paper aims to value data by asking the following question: How much do data processing activities contribute to a firm's total revenues? Data processing here refers broadly to nearly all kinds of data-related activities, including data collection, data cleansing, data analysis, and data storage. A simple OLS model regressing sales on data processing activities, however, would encounter two problems. First, this naive approach suffers from endogeneity issues. The relation may run from sales to data instead. Firms with high sales may possess more resources to collect and analyze more data, reversing the direction of causality. Or, a hidden factor could influence both data and sales, biasing estimates. Second, without access to firms' internal records, it is hard to measure the quantity of their data processing activities or the types of data they process.

To address the first challenge of endogeneity, I leverage the exogenous variation in data processing activities introduced by the General Data Protection Regulation (GDPR). Enacted by the European Union (EU), the GDPR is a comprehensive law designed to safeguard the privacy and personal data of EU residents. It establishes rules governing how firms collect, analyze, and handle personal data, imposing substantial fines of 4% of global sales for non-compliance, irrespective of the firms' locations. Within a difference-in-differences (DiD) framework, I assign firms with larger legal departments to the treatment group. Survey evidence suggests that legal departments often play a supportive and leading role in ensuring GDPR compliance ([European Data Protection](#)

Board, 2024; European Center for Digital Rights, 2024). Powerful legal departments would thus compel their respective firms to reduce data processing activities post-GDPR to comply with this data protection law, rather than assisting their firms in navigating lawsuits while maintaining high levels of data processing. My findings indicate that treated firms experience simultaneous declines in data processing and sales after GDPR's implementation, confirming the contribution of data to revenues.

To address the second challenge of measuring data processing activities, I focus on an observable component: third-party web tracking. Firms deploy website technologies to track user behavior. Counting the number of web technologies used thus provides a measurable, albeit incomplete, picture of overall data processing. Since GDPR impacts all kinds of data processing, I expect web technology adoption to co-vary with the level of overall data processing after GDPR's implementation. Therefore, I utilize the inverse hyperbolic sine of the web technologies used as a proxy for overall data processing activities.

The sample consists of 2,371 U.S. public firms observed over a seven-year period bracketing the implementation of GDPR in 2018. I hypothesize that firms with larger legal departments will be more likely to limit data processing after GDPR. The treatment group consists of firms with above-median legal staff shares. DiD regressions indicate treated firms reduce data processing activities (proxied by web technologies used) by 11% relative to control firms after GDPR's implementation in May 2018. Legal staff shares should not affect the future growth of data processing or sales absent significant privacy regulations like GDPR, satisfying DiD's parallel trend assumption. I also find parallel pre-trends that indicate the parallel trend assumption is likely to be true. The observed GDPR treatment effect on data processing activities thus likely extracts the exogenous reduction in data processing activities. This effect also concentrates on analytics web technologies, which are GDPR's primary targets.

I next conduct DiD analysis regressing firm fundamentals on the treatment indicator. Treated firms exhibit a 1.2% decline in sales (scaled by 2016Q1 assets) relative to control firms post-GDPR, while the trends for income before extraordinary items and net income

remain parallel. SG&A declines for treated firms, but the CAPX trend remains parallel. All tested fundamentals exhibit parallel pre-trends. For the median firm with \$1.1 billion of assets in 2016Q1, this 1.2% sales decline translates to \$13.2 million quarterly revenue losses in the post-period.

If data processing is the main channel from GDPR to sales, the 11% data processing decline induced by GDPR, combined with the 1.2% sales decline, implies a 1% increase in data processing activities can raise sales by 0.11%. This establishes data's value by answering the question of how much data processing contributes to a firm's total revenues. One caveat is that GDPR can impact a firm in two main ways: affecting data processing activities and imposing compliance costs. Compliance costs, rather than reduced data processing, could drive the sales decline. However, compliance costs do not directly enter into sales calculation. Additionally, treated firms do not experience a rise in SG&A expenditures relative to control firms after GDPR. The two facts suggest compliance costs likely do not mediate the relationship between GDPR and sales declines, strengthening the inference that reduced data processing, not competing channels, drives falling revenues.

Instrumental variable (IV) regressions, where treatment times post is used to instrument for data processing activities, also yield similar results. These regressions suggest that a 1% increase in data processing can raise sales by 0.17%, which is close to the 0.11% estimate found in the previous paragraph. Therefore, the IV regressions support the finding that the GDPR-induced decline in data processing activities substantially contributes to the observed sales decline.

After establishing the relation from data processing activities to sales, I investigate heterogeneity in the impacts of GDPR. Among treated firms, those with lower shares of software engineers or workers located in EU exhibit greater data processing declines post-GDPR. Firms that employ a large proportion of software engineers, thus possessing software expertise, can more easily devise new methods that process data in compliant ways, limiting the data processing drop to only 5%. Firms lacking software expertise reduce data processing by 17% because they have few options beyond halting data pro-

cessing altogether if they want to comply with GDPR. On the EU worker split, firms minimally exposed to the EU face the same 4% fine on global revenues but derive little reward from EU, so they avoid these low-premium risks by cutting data processing. These two groups that cut more data processing post-GDPR also see larger sales declines. They even show significant declines in income before extraordinary items and net income, though such declines are insignificant when I pool all treated firms in the main results. This further supports the inference that data processing contributes to revenues.

This paper contributes to two strands of literature. First, it adds to the economic research on valuing data. As summarized by [Veldkamp \(2023\)](#), attaching value to data in economic contexts is challenging since most data are byproducts of business activities and thus lack explicit price tags. Existing studies typically infer data's value through firm outcomes induced by data usage. Researchers have adopted structural models to quantify the value of data. For instance, [Abis and Veldkamp \(2022\)](#) incorporate data into production functions, [Eeckhout and Veldkamp \(2022\)](#) analyze the covariance between production and profit margin, and [Farboodi et al. \(2022\)](#) clear equilibrium asset pricing models that incorporate data. In marketing literature, researchers are also interested in privacy regulations and their impacts on advertisements. They find that customer data access enables more effective targeted advertising, which is also one form of data's value. They analyze, e.g., the EU Privacy Directive's impact on the advertising industry ([Goldfarb and Tucker, 2011](#)), or GDPR's impact on a telecommunications company ([Godinho de Matos and Adjerid, 2022](#)). Machine learning adoption, a form of data use, is also found to increase sales, market value, number of employees ([Babina et al., 2021](#); [Rock, 2019](#)), and demand for skilled labor ([Babina et al., 2022](#)). I contribute to this literature by using the GDPR, a regulatory shock, as a natural experiment to analyze data activities. I link the exogenous variation in data processing to changes in financial performance. Furthermore, I apply this setting to public firms in the U.S. in general, extending the analysis beyond specific industries examined in prior studies.

Second, this paper adds to the growing economic literature on GDPR and privacy regulations in general. Prior work finds rising concentration in web technology markets

post-GDPR, as firms favor web technologies provided by large vendors, who are perceived as more GDPR-compliant (Johnson et al., 2023). In the short run, third-party online interactions decline (Peukert et al., 2022) along with firm performance and app market entry (Chen et al., 2022; Janßen et al., 2022). Johnson (2022) provides an excellent review of the economic literature on GDPR. Other privacy regulation shocks are also shown to impact economic activities. For example, Apple app store privacy policies hurt app downloads (Bian et al., 2021), and AdChoices opt-out program results in wasted advertising expenses (Johnson et al., 2020). I differentiate my approach from this existing literature in three key ways. First, I aim to quantify data’s value by using GDPR as a shock, rather than studying GDPR itself. Second, most empirical GDPR research relies on high-frequency data over a short period around implementation, while I examine a long seven-year period. Third, existing papers use less robust treatment groups like EU versus U.S. firms. I employ ex-ante size of legal departments within firms, measured by the share of legal staff, as a more exogenous allocation variable.

The remainder of the paper proceeds as follows. Section 2 outlines background information on GDPR, the treatment group design, and web tracking. Section 3 examines GDPR’s impact on data processing activities and firm fundamentals, providing evidence that data processing increases sales. Section 4 explores heterogeneous treatment effects of GDPR across subgroups. Section 5 conducts robustness checks. Finally, Section 6 summarizes the findings.

2 Background Information and Data

In this section, I explain how the General Data Protection Regulation (GDPR) affects data processing and how it inspires the treatment and control group design in the difference-in-differences (DiD) analysis. I also describe how I measure data processing activities by observing a narrow part of it, web tracking. Finally, I present the data sources and the sample construction process.

2.1 General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) is a regulation introduced by the European Union (EU) to enhance the protection of personal data privacy. The EU announced it on April 14, 2016, and implemented it on May 25, 2018. The GDPR safeguards EU residents' personal data, such as names, locations, health records, and online identifiers (e.g., IP addresses and cookies) (Article 4). It also sets principles for data processing (Article 5). For example, firms can process only the data that are necessary and relevant to pre-specified purposes (data minimization), and store the data only for the necessary duration (storage limitation). In practice, firms may delete user data after say one year.

Data processing must also be lawful, meeting at least one of the six legal bases: consent, contractual obligation, legal obligation, vital interest, public interest, or legitimate interest (Article 6). For instance, firms can lawfully process customer data after acquiring explicit consent. They can also process data that are required to fulfill contractual obligations, like processing names and addresses in order to deliver ordered products. GDPR can fine violators up to 4% of annual global revenues, and this applies even to non-EU firms that handle EU customers' data (Article 83). Therefore, firms around the globe, including U.S. firms, would reduce data processing activities after GDPR's implementation.

2.2 DiD Treatment Group Design

U.S. firms that reduce more data processing activities after GDPR thus form my treatment group, and these are firms with larger (more powerful) legal departments. According to [European Data Protection Board \(2024\)](#), an investigation into data protection officers reveals that most of them work in legal departments. Survey evidence from [European Center for Digital Rights \(2024\)](#) also suggests that legal departments generally support GDPR compliance, rather than assisting the firms in navigating lawsuits while maintaining high levels of data processing. Thus, I hypothesize that when a firm's legal department is more powerful, it will compel the firm to reduce data-related activities when significant privacy regulations, such as GDPR, are introduced.

I define the treatment indicator *High Legal* as one if the share of employees working as legal staff in March 2016 is above the median. In the absence of major regulations like GDPR, the size of the legal department should not influence future data processing or financial performance, in the same way that firefighters do not influence a city’s development when there is no fire. Both treated and control firms should exhibit parallel growth trends in the absence of GDPR, satisfying DiD’s parallel trend assumption. I also find parallel pre-trends in the data that support this assumption.

To measure the number of legal staff and total employees, I obtain resume data from Revelio Labs, a workforce intelligence company that structures millions of unstructured public employment records into a firm-position dataset. Unlike job posting datasets such as Burning Glass Technologies which measure a firm’s demand for skills, the resume dataset records employees already onboard, measuring a firm’s existing capabilities or characteristics. In addition to job titles and employment periods, Revelio Labs also designates individual jobs into functional roles like legal, software, or business development. I calculate the legal department’s size by dividing a firm’s legal workers in March 2016 by its total employees. I set the indicator variable *High Legal* to be one if this ratio is above the median. In heterogeneity analysis, I also define variables in similar ways indicating low ratio of software engineers (*Low SW*) and low ratio of workers located in EU (*Low EU*).

A closer look into the composition of treated firms reveals that the industry with the highest share of treated firms is the financial industry (2-digit NAICS 52), with 93% of firms being treated. However, in the March 2016 cross-section, being a *High Legal* firm is uncorrelated with being a *Low SW* (software) firm or a *Low EU* firm, with correlations of 0.12 and -0.02 , respectively. Firms with large legal departments do not seem to have higher exposure to EU or hire more software engineers. Table A.1 further compares the treatment and control groups by regressing the *High Legal* indicator on fundamentals in the 2016Q1 cross-section. Treated firms likely differ, being larger and less profitable with lower SG&A spending. But as long as data processing and financial performance dynamics are parallel absent GDPR, the DiD inference remains valid.

2.3 Third-party Web Tracking and Website Technologies

Having defined the treatment group, the next step is measuring the level of data processing. Precisely gauging a firm's data usage is difficult without access to internal records. However, web tracking, a subset of data processing, can be more easily observed and quantified. An example of web tracking would be that if people search for a product on Google, they may later see related Facebook advertisements. Facebook tracks identities and analyzes browsing history to determine users' interests. Figure 1 illustrates the relationship between overall data processing activities and web tracking before and after GDPR. While web tracking captures only a fraction of overall data processing, it is readily measurable. When significant privacy regulations such as GDPR hit, overall data processing and web tracking should shrink together. Thus, observed changes in tracking intensity reveal shifts of unobserved overall data processing efforts in the same direction.

Websites install web technologies to track user behavior. Third-party vendors provide these technologies, allowing websites to perform additional functionalities. For example, my university's official website installs Google Analytics for visitor tracking, reCAPTCHA for bot protection, Facebook for social media engagement, and YouTube for video streaming. Most web technologies are related to web tracking and data processing. For example, analytics technologies like Google Analytics explicitly track and analyze user behavior. Even non-analytics technologies such as social media technologies may also facilitate tracking and data processing, for they must locate and identify users before interacting with them.

As most web technologies are related to web tracking, I count a firm's time-varying web technology usage as the measure for web tracking, which serves as a proxy for overall data processing activities. I obtain website-technology pairs from BuiltWith, a technology company specializing in website profiling. BuiltWith offers tools such as browser plugins that enable users to identify the technologies loaded on websites they visit. BuiltWith also compiles a historical database of website-technology pairs with adoption and removal dates by periodically scanning popular websites. I am able to match BuiltWith data to 4,013 U.S. public firms in Compustat.

Specifically, I proxy data processing activities with the inverse hyperbolic sine, \sinh^{-1} , of the number of web technologies that a firm uses in a month. BuiltWith also classifies technologies into functional groups. I count total technologies and analytics technologies separately in some analyses, as analytics technologies explicitly process user data for insights and are GDPR’s primary targets.

2.4 Sample Selection and Other Datasets

The sample consists of U.S. public firms from 2014 to 2020, a span of seven years, starting from two years before GDPR’s announcement (April 2016) to two years after GDPR’s implementation (May 2018). I aggregate BuiltWith’s web technology adoption into firm-month level observations. Firm fundamentals such as sales come from Compustat Quarterly at firm-quarter level. In most tests, I allocate firms into groups based on their ex-ante traits in March 2016 or 2016Q1, right before GDPR’s announcement. After excluding firms with missing web technology, employee records, or ex-ante traits, I arrive at a sample of 2,371 firms.

I winsor all continuous variables at 2% and 98% levels. Table 1 reports the summary statistics for variables used. All regressions report standard errors clustered at the firm level. As the web technology data are still under-studied in the literature, Table 2 offers a diagnostic analysis by regressing web technology adoption on firm fundamentals, providing context on potential determinants of web technology adoption. Column 1 shows total web technology adoption’s correlations with firm fundamentals, Column 2 shows analytics technologies, and Column 3 shows non-analytics technologies. Regressions show that technology adoption is orthogonal to most fundamentals. Only firm size exhibits a significant correlation, which indicates that firms of different sizes may follow distinct data processing dynamics. Thus, nearly all subsequent regressions include both firm and time-by-2016Q1 asset quintile fixed effects, addressing both time-invariant traits and time-varying dynamics in each of the five size groups.

3 Data Processing Activities' Contribution to Revenues

This section investigates GDPR's impacts on data processing activities and financial performance using DiD analysis. The results support the inference that data contributes to total revenues. This section ends by re-interpreting the same findings within an instrumental variable (IV) framework.

3.1 Data Processing Activities Over Time

This section investigates data processing activities over time for the treatment group in a dynamic DiD framework. Treated firms are those with more powerful legal departments, defined as having above-median legal staff shares in March 2016. As discussed in Section 2.2, I hypothesize that treated firms are forced by their legal departments to restrict data processing post-GDPR. The regression formula is

$$\begin{aligned} IHS(No. Tech_{i,t}) = & \sum_{s \neq Mar2016} \beta_s \times \mathbb{1}(s = t) \times High Legal_i \\ & + Firm FE + Month-Asset Quintile FE + \epsilon_{i,t}. \end{aligned}$$

The dependent variable measures data processing activities, being inverse hyperbolic sine of number of web technologies used by firm i in month t . $High Legal_i$ is the time-invariant treatment indicator being one if the share of legal staff is above the median in March 2016. There is one beta for each month, and the left-out month is March 2016, one month before GDPR's announcement. Firm fixed effects rule out invariant firm characteristics. As discussed in Section 2.4, firms of different sizes might experience different dynamics in data processing activities, so I also include month-by-2016Q1 asset quintile fixed effects.

Figure 2 plots the monthly coefficients with 90% confidence intervals. The red dots correspond to total web technologies, the main data processing measure. After GDPR's announcement, treated firms start reducing data processing activities compared with control firms. After GDPR's actual implementation, the data processing activities of

treated firms reduce even more, significantly diverging from control firms by 10-15%. In the pre-GDPR years, data processing trends remain roughly parallel between groups.

To provide more context on data processing dynamics, I offer two supplementary analyses. First, the same Figure 2 also analyzes analytics technologies and plots coefficients in blue squares. Web technologies are categorized into functional groups by BuiltWith, such as advertising, analytics, e-commerce, or social media widgets. Among the categories, analytics technologies explicitly process data, so they are GDPR’s primary targets. Analytics technologies decline substantially more than average technologies, with most blue squares falling outside the red error bands post-GDPR. Even within web tracking, some activities are more affected by GDPR, and they indeed show greater declines after GDPR’s implementation.

Second, Figure 3 plots absolute data processing levels over time separately for treatment and control groups, providing a more complete picture of the trends. It graphs monthly fixed effects for the control group, and monthly fixed effects plus beta with 90% confidence intervals for the treated group. In the pre-GDPR era, both groups exhibit similar growth in data processing. After implementation, the treated group still increases processing but at a significantly slower pace compared with control firms.

3.2 Firm Fundamentals Over Time

After establishing the relation between GDPR and data processing activities, this section investigates firm fundamentals for the treatment group over time in a dynamic DiD framework. The statistical model is

$$\begin{aligned}
 Dep\ Var_{i,t+1} = & \sum_{s \neq 2016Q1} \beta_s \times \mathbb{1}(s = t) \times High\ Legal_i \\
 & + Firm\ FE + Month - Asset\ Quintile\ FE + \epsilon_{i,t}.
 \end{aligned}$$

$Dep\ Var_{i,t+1}$ is firm i ’s fundamentals in quarter $t + 1$. There is one beta for each quarter, and the left-out quarter is 2016Q1, one quarter before GDPR’s announcement. This model closely follows Section 3.1 but uses quarterly data. All dependent variables, like

sales, are scaled by firm i 's assets in 2016Q1. [Welch \(2021\)](#) warns that commonly adopted time-varying scaling variables, such as contemporary assets, tend to confound with other variables and introduce spurious correlations in regression models. Thus, I scale all fundamentals on the left-hand side by the time-invariant 2016Q1 assets.

Panel [A](#) of [Figure 4](#) plots dynamic DiD coefficients with dependant variables being sales, income before extraordinary items (IB), and net income (NI), all scaled by 2016Q1 assets. Sales show a significant decline for the treatment group relative to the control group after GDPR's announcement and implementation, with the magnitude being around 1–2% of 2016Q1 assets. The upward spike in sales around 2020Q2 is likely due to COVID-19, which disrupted business plans broadly in 2020. Income before extraordinary items and net income exhibit no declines, and I will comment on this finding at the end of this subsection.

If the decline in data processing activities is the sole or at least a significant channel between GDPR and sales decline, then the observed reduction in data processing is the factor that drives the decline in sales, implying a positive relationship from data to sales. However, GDPR can impact firms in two main ways. Beyond restricting data processing, it may also impose substantial compliance costs. In some industries, these costs could even exceed GDPR's maximum fines ([Johnson, 2022](#)). Lawsuits due to GDPR may also lead to financial costs for treated firms. Reassuringly, compliance costs do not directly factor into sales calculations, so the sales results are unlikely confounded by GDPR-imposed costs. Additionally, in Panel [B](#) of [Figure 4](#), I plot the same dynamic DiD model but with dependant variables changed to expenses such as CAPX and SG&A. Treated firms experience lower SG&A, suggesting no relative compliance cost surge. Together, these facts indicate that reduced data processing, rather than competing channels like compliance costs, is the more likely channel between GDPR and sales declines.

I would like to conclude this subsection with a discussion on the flat dynamics of income measures in Panel [A](#) of [Figure 4](#), namely net income (NI) and income before extraordinary items (IB). Although these flat income measures may seem to undermine my conclusion that data processing brings value to the firm, I interpret the income measures

in two ways. First, the reduction in sales and reduction in SG&A may cancel each other out, leading to no observable changes in the final income variables over time. Second, in Section 4, I find indirect evidence that a reduction in data processing leads to a reduction in income variables when I further split the treated firms into subgroups. Therefore, it is likely that data processing contributes to net income and income before extraordinary items, but such effects are obscured when pooling all treated firms together. In my main analysis, however, I focus on sales as a cleaner measure of a firm’s business performance, without being confounded by costs, taxes, and interest rates.

3.3 Two-period Difference-in-differences Specification

This section repeats the analysis in Sections 3.1 and 3.2, but in a two-period DiD framework. They produce compact regression tables, and pave the way for clearer instrumental variable regressions and heterogeneity analysis in later sections.

To transform the dynamic DiD regressions into two-period DiD, I use the following regression model

$$Dep\ Var_{i,t} = \beta_1\ High\ Legal_i \times Post_t + Firm\ FE + Time\text{-}Asset\ Quintile\ FE + \epsilon_{i,t}.$$

$Dep\ Var_{i,t}$ for firm i is either the inverse hyperbolic sine of the number of web technologies in month t , or firm fundamentals scaled by 2016Q1 assets in quarter $t+1$. $High\ Legal_i$ is a time-invariant treatment indicator being one if the share of legal staff is above the median in March 2016, and zero otherwise. $Post_t$ is one after May 2018, GDPR’s implementation, and zero otherwise.¹

Table 3 tabulates GDPR’s impact on data processing and fundamentals for treated firms. Column 1 reflects Figure 2, reiterating *High Legal* firms’ larger data processing decline post-GDPR. Treated firms drop 11% in data processing activities relative to control firms post-GDPR. Other columns reflect Figure 4, showing decreases of sales by 1.2% of 2016Q1 assets, and SG&A by 0.23%, while income before extraordinary items,

¹I also define $Post_t$ alternatively by GDPR’s announcement in April 2016 in Section 5, finding qualitatively similar results.

net income, and CAPX remain in parallel trends in the post-period. The 1.2% sales drop translates to \$13.2 million quarterly revenue losses for the median firms with \$1.1 billion assets in 2016Q1. If the 11% data processing decline contributes fully to the 1.2% sales decline, it indicates 1% increase in data processing activities can drive 0.11% increase in sales. I omit controls here since Table 2 indicates web technology adoption is largely orthogonal to firm fundamentals. Nevertheless, Table A.2 extends Table 3 by including an extensive set of control variables. It shows identical results — declines for treated firms in technologies, sales, and SG&A, but parallel trends for income variables and CAPX. However, many firm controls are highly correlated with the dependent variables, for example SG&A can appear on both left and right-hand sides. For clarity and to maintain the sample size, I exclude control variables in subsequent regressions.

3.4 Two-stage Least Squares Regression

Table 4 strengthens the evidence for data's effect on sales using instrumental variables (IV). Column 1 is a naive simple OLS model regressing sales on data, showing a positive correlation between data and sales. Correlation does not reveal the value of data, because data processing may also be affected by sales if profitable firms can afford to collect and analyze more data, or a third factor could drive both data and sales. After the previous discussions, however, *High Legal* times *Post* becomes a natural candidate as the IV for the endogenous data processing activities. Column 2 shows the first stage, regressing data processing on the IV. The IV negatively predicts data processing which satisfies the relevance condition, with the F-statistic of 10.4 passing the conventional threshold of 10. Column 3 shows the second stage with a significantly positive effect of data on sales. The IV coefficient suggests 1% higher data processing drives sales up 0.17%, similar in magnitude to the 0.11% estimate inferred from the previous subsection. The IV estimate is also around 20 times the OLS estimate, implying OLS underestimates the true effect. It is worth noting that inflated IV estimates may exaggerate the true magnitude (Jiang, 2017). Still, the IV estimate here should suggest a sizable effect of data on sales.

4 Heterogeneity in GDPR's Treatment Effects

After establishing core findings in the previous section, this section explores the heterogeneous treatment effects of GDPR across firms of different traits. This analysis dives deeper into the second-order patterns of GDPR's impacts. The analysis also serves as indirect evidence of data's contribution to sales or even incomes.

4.1 Share of Software Workforce

Firms are forced by their legal departments to adjust data processing under the GDPR. But even among treated firms, they may differ in their abilities to adjust data-related practices. This section examines software expertise, measured by March 2016 software engineer share, as a required adjustment ability.

Panel A of Table 5 interacts $High\ Legal \times Post$ with $Low\ SW$, an indicator equal to one if a firm's software engineer share is below-median. Column 1 shows that among treated firms, those with high software workforce reduce data processing by a marginally significant 5% post-GDPR. On the other hand, firms with low software workforce reduce processing by a substantially higher 17% ($0.0501 + 0.1193$). Software expertise may enable firms to discover new data processing methods that are compliant, without reducing data processing. While firms with little software expertise have to cut data processing to stay compliant. Column 2 shows declines in sales concentrate in the same $Low\ SW$ group that cut more data processing, indirectly suggesting that it is the data processing decline that drives the reduction in sales.

Columns 3 and 4 find decreases in income before extraordinary items and net income for treated firms with low software expertise. Previously when I pool all treated firms in Table 3, income variables exhibit no decline. But the subgroup responsible for the data processing reduction does show income decreases, indirectly evidencing that data may positively drive incomes too. Still, this paper focuses on sales over income variables, because incomes are potentially confounded by costs, taxes, and interest rates. Column 5 shows the $Low\ SW$ group also exhibits more SG&A declines.

4.2 Share of EU Workforce

Panel **B** of Table 5 examines heterogeneity by the exposure to EU, defined as the share of employees located in EU. I define the firm-level *Low EU* indicator as one if the share of employees located in EU is above median in March 2016. Column 1 shows among *High Legal* firms, *Low EU* firms drop more data processing post-GDPR. This aligns with findings in [Johnson et al. \(2023\)](#), that some firms derive lower benefits from EU yet still risk the same 4% fine on global revenues. These *Low EU* firms face imbalanced risks and rewards, and the safest way to shed the low-premium risks is by reducing data processing rather than processing data in a compliant way. Column 2 shows sales declines concentrate in the same *Low EU* firms, indirectly supporting data's contribution to revenues. Columns 3 and 4 display income decreases, and Column 5 shows lower SG&A, all for the *Low EU* firms.

5 Robustness Checks

This section explores four robustness tests. The first test addresses a 2018 privacy law in California, the California Consumer Privacy Act (CCPA). The other three test the sensitivity of the results to different assumptions. California implemented CCPA in 2018, the same year as GDPR's implementation. A possible concern is that CCPA, rather than GDPR, drives my results. I would like to emphasize that this is not a paper focusing on GDPR; I use GDPR merely as a tool to shock data processing activities. If CCPA can also affect data processing in treated firms, my results of data's effect on sales would still hold. Nevertheless, to rule out the confounding effects of CCPA, I exclude California firms from the sample and re-estimate the regressions. Panel **A** of Table 6, which extends from Table 3 and drops California firms, shows that the coefficients on data processing and sales remain negative, while SG&A remains non-increasing. Excluding California firms does not qualitatively change the results.

The rest of the tables test the robustness of the results to different definitions of the post-treatment period and outlier removal, as well as to the exclusion of the COVID-

19 crisis. Previous regressions define $Post_t$ by GDPR’s implementation date, but firms may adjust data processing preemptively before implementation. Panel **B** defines $Post$ alternatively by GDPR’s announcement date. Panel **C** purges outliers more aggressively, changing the winsor levels to 5% and 95%. This ensures that the results are not driven by outliers. Panel **D** excludes observations in the year 2020 from the sample to rule out the effect of the COVID-19 crisis, which disrupted business activities in the last year of the sample.² All three tests indicate that changing the assumptions does not qualitatively alter the results.

6 Summary

This paper identifies the positive value of data to U.S. public firms. The DiD framework uses web tracking as a proxy for data processing activities, firms with large legal departments as the treatment group, and the GDPR as an exogenous shock to data processing activities. I find that the treated firms reduce data processing by 11% and experience a 1.2% decline in sales relative to the control firms post-GDPR. This result suggests that the exogenous decline in data processing drives the decline in sales, establishing the positive value of data. Additional evidence, such as treated firms’ decline in SG&A, further supports the data processing channel and rules out alternative stories such as the compliance risks channel. Using $Treat$ times $Post$ to instrument for data processing activities, I also find significant second stage coefficients indicating that data processing activities have a positive causal effect on sales. Heterogeneity analysis reveals that firms with low software expertise and few EU workers are primarily responsible for the observed drop in data processing among treated firms. These same subgroups of treated firms are also responsible for the decline in sales, indirectly evidencing that GDPR affects financial performance through the data processing channel.

²Such disruption is visible in sales dynamics in Panel **A** of Figure 4.

References

- Abis, Simona, and Laura Veldkamp, 2022, The Changing Economics of Knowledge Production, *Working Paper* .
- Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, 2021, Artificial Intelligence, Firm Growth, and Product Innovation, *Working Paper* .
- Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, 2022, Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition, *Working Paper* .
- Bian, Bo, Xinchun Ma, and Huan Tang, 2021, The Supply and Demand for Data Privacy: Evidence from Mobile Apps, *Working Paper* .
- Chen, Chinchih, Carl Benedikt Frey, and Giorgio Presidente, 2022, Privacy Regulation and Firm Performance: Estimating the GDPR Effect Globally, *Working Paper* .
- Eeckhout, Jan, and Laura Veldkamp, 2022, Data and Market Power, *Working Paper* .
- European Center for Digital Rights, 2024, GDPR: A culture of non-compliance?, Technical report.
- European Data Protection Board, 2024, Designation and Position of Data Protection Officers, Technical report.
- Farboodi, Maryam, Dhruv Singal, Laura Veldkamp, and Venky Venkateswaran, 2022, Valuing Financial Data, *Working Paper* .
- Godinho de Matos, Miguel, and Idris Adjerid, 2022, Consumer Consent and Firm Targeting After GDPR: The Case of a Large Telecom Provider, *Management Science* 68, 3330–3378.
- Goldfarb, Avi, and Catherine E. Tucker, 2011, Privacy Regulation and Online Advertising, *Management Science* 57, 57–71.

- Janßen, Rebecca, Reinhold Kesler, Michael E. Kummer, and Joel Waldfogel, 2022, GDPR and the Lost Generation of Innovative Apps, *Working Paper* .
- Jiang, Wei, 2017, Have Instrumental Variables Brought Us Closer to the Truth, *Review of Corporate Finance Studies* 6, 127–140.
- Johnson, Garrett, 2022, Economic Research on Privacy Regulation: Lessons From the GDPR and Beyond, *Working Paper* .
- Johnson, Garrett A., Scott K. Shriver, and Shaoyin Du, 2020, Consumer Privacy Choice in Online Advertising: Who Optes Out and at What Cost to Industry?, *Marketing Science* 39, 33–51.
- Johnson, Garrett A., Scott K. Shriver, and Samuel G. Goldberg, 2023, Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR, *Management Science* 69, 5695–5721.
- Peters, Ryan H., and Lucian A. Taylor, 2017, Intangible capital and the investment-q relation, *Journal of Financial Economics* 123, 251–272.
- Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer, 2022, Regulatory Spillovers and Data Governance: Evidence from the GDPR, *Marketing Science* 41, 746–768.
- Rock, Daniel, 2019, Engineering Value: The Returns to Technological Talent and Investments in Artificial Intelligence, *Working Paper* .
- Veldkamp, Laura, 2023, Valuing Data as an Asset, *Review of Finance* 27, 1545–1562.
- Welch, Ivo, 2021, Spurious Inference Caused by Time-Series Variation in Scaling: Real Estate Shocks Did Not Affect Corporate Investment, *Working Paper* .

Figure 1: Relationship between Overall Data Processing and Web Tracking

Web tracking is a subset of data processing activities. After GDPR, I hypothesize that web tracking and data processing activities of affected firms both decline. As a result, the easily measurable web tracking serves as a proxy for the level of overall data processing activities

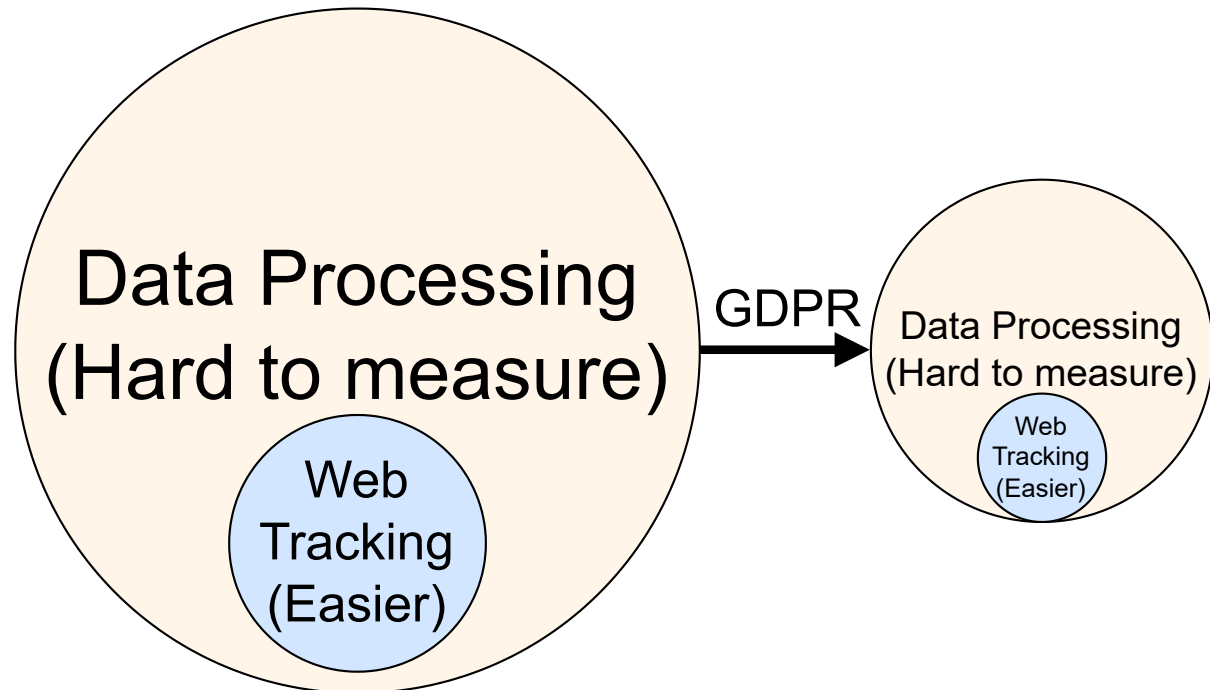


Figure 2: Data Processing Activities Over Time

This figure plots coefficients and 90% confidence intervals from dynamic difference-in-differences regressions examining the impact of GDPR on firms' data processing activities. The regression model estimated on firm-month panel data is

$$IHS(No. Tech_{i,t}) = \sum_{s \neq Mar2016} \beta_s \times \mathbb{1}(s = t) \times High Legal_i + Firm FE + Month-Asset Quintile FE + \epsilon_{i,t}.$$

The dependent variable measures data processing activities, being the inverse hyperbolic sine of the number of web technologies used by firm i in month t , including all technologies as well as solely analytics technologies. $High Legal_i$ is a time-invariant indicator equal to one if firm i 's share of legal staff is above the median in March 2016. There is one beta for each month, and the left-out month is March 2016. The regressions include firm and month-by-2016Q1 asset quintile fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%. Vertical dashed lines mark March 2016, right before GDPR's announcement, and April 2018, right before GDPR's implementation.

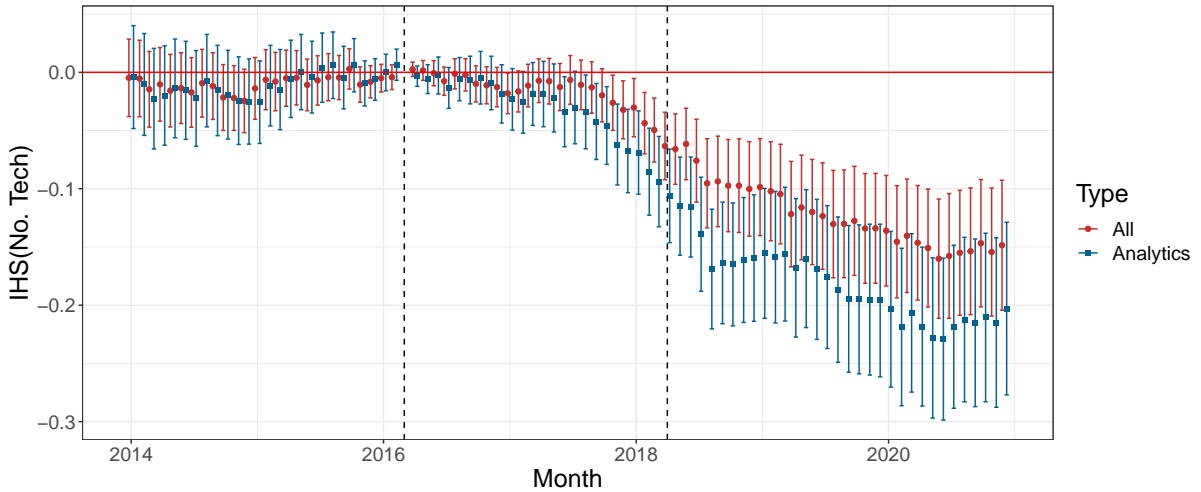


Figure 3: Data Processing Activities Over Time: Plotting both Treatment and Control

This figure plots beta plus month fixed effects (treatment) and month fixed effects alone (control) with 90% confidence intervals from dynamic difference-in-differences regressions examining the impact of GDPR on firms' data processing activities. The regression model estimated on firm-month panel data is

$$IHS(No. Tech_{i,t}) = \sum_{s \neq Mar2016} \beta_s \times \mathbb{1}(s = t) \times High Legal_i + Firm FE + Month FE + \epsilon_{i,t}.$$

The dependent variable measures data processing activities, being the inverse hyperbolic sine of the number of web technologies used by firm i in month t , including all technologies as well as solely analytics technologies. $High Legal_i$ is a time-invariant indicator equal to one if firm i 's share of legal staff is above the median in March 2016. There is one beta for each month, and the left-out month is March 2016. The regressions include firm and month fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%. Vertical dashed lines mark March 2016, right before GDPR's announcement, and April 2018, right before GDPR's implementation.

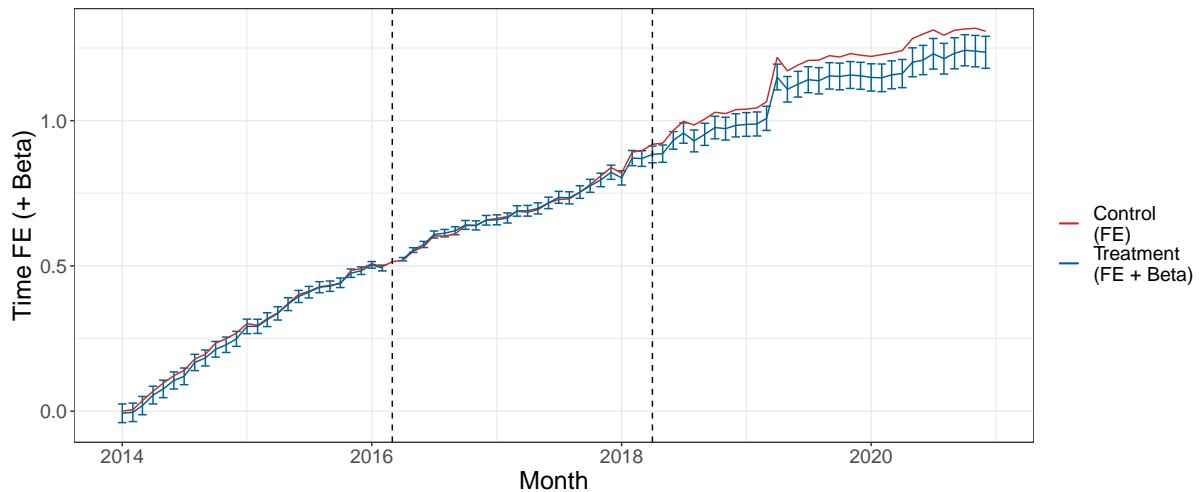


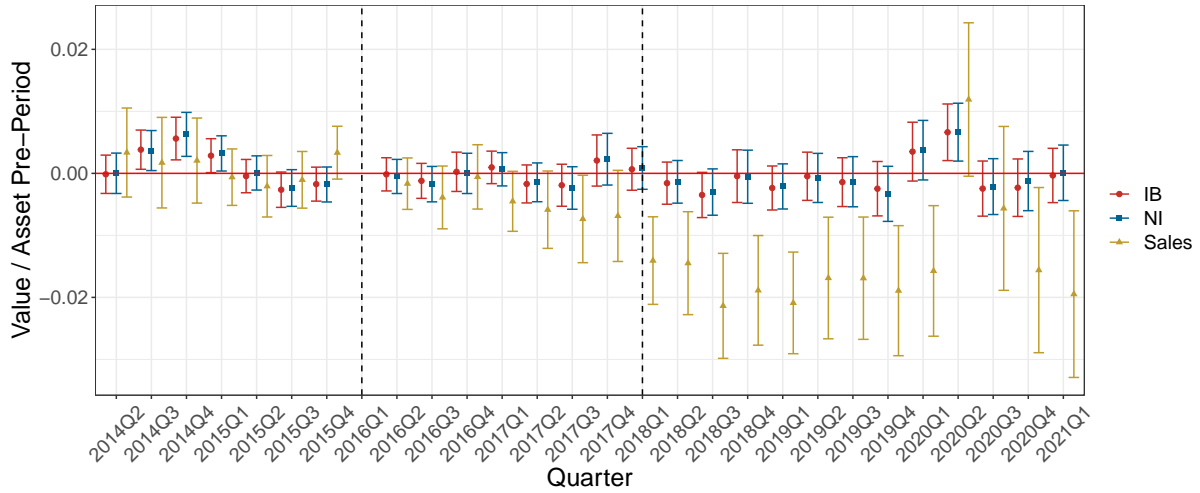
Figure 4: **Firm Fundamentals Over Time**

This figure plots coefficients and 90% confidence intervals from dynamic difference-in-differences regressions examining the impact of GDPR on firms' fundamentals. The regression model estimated on firm-quarter panel data is

$$\begin{aligned}
 Dep\ Var_{i,t+1} = & \sum_{s \neq 2016Q1} \beta_s \times \mathbb{1}(s = t) \times High\ Legal_i \\
 & + Firm\ FE + Month\text{-}Asset\ Quintile\ FE + \epsilon_{i,t}.
 \end{aligned}$$

The dependent variable is firm characteristics for firm i in quarter $t + 1$ scaled by 2016Q1 assets. $High\ Legal_i$ is a time-invariant indicator equal to one if firm i 's share of legal staff is above the median in March 2016. There is one beta for each quarter, and the left-out quarter is 2016Q1. Panel **A** plots sales, income before extraordinary items (IB), and net income (NI) as dependent variables. Panel **B** plots CAPX and SG&A as dependent variables. The regressions include firm and quarter-by-2016Q1 asset quintile fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%. Vertical dashed lines mark 2016Q1, right before GDPR's announcement, and 2018Q1, right before GDPR's implementation.

(A) **Revenues and Profits**



(B) **Expenses**

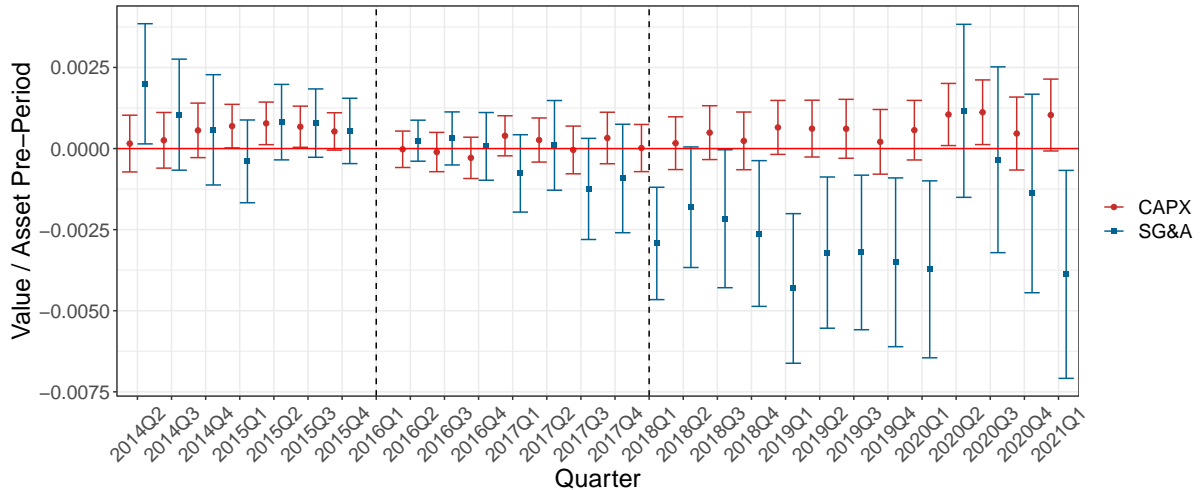


Table 1: **Summary Statistics**

This table provides an overview for the variables used in this paper. Sorting variables are firm-level characteristics measured in March 2016 or 2016Q1. Dependent variables are scaled by 2016Q1 assets. Control variables that are scaled by assets use lagged contemporary assets as the scaling variable.

	N	Mean	Std. Dev.	Min	P25	Median	P75	Max
Data Processing								
IHS(No. Tech)	184,520	4.609	0.862	2.644	4.061	4.605	5.182	6.519
IHS(No. Tech ^{Analytics})	184,520	2.231	1.057	0	1.444	2.095	2.998	4.304
Sorting Var								
Share Legal	184,520	0.013	0.015	0	0.003	0.007	0.017	0.067
Share SW	184,520	0.049	0.072	0	0.007	0.017	0.052	0.308
Share EU	184,520	0.088	0.097	0	0.015	0.045	0.139	0.379
KZ Index	157,147	0.957	1.509	-3.25	0.132	0.925	1.749	5.769
Dep Var								
Sales	54,960	0.236	0.223	0.004	0.065	0.179	0.332	0.982
IB	54,989	-0.005	0.056	-0.239	-0.007	0.005	0.019	0.088
NI	54,983	-0.005	0.057	-0.241	-0.007	0.005	0.019	0.091
SG&A	55,042	0.05	0.062	0	0.005	0.029	0.071	0.269
CAPX	54,502	0.01	0.013	0	0.001	0.005	0.013	0.064
Controls								
Size	55,253	7.122	2.164	2.433	5.68	7.216	8.603	11.644
Cash Flow	51,689	0.003	0.051	-0.205	0.001	0.014	0.028	0.082
ROA	55,067	-0.006	0.05	-0.211	-0.006	0.005	0.017	0.07
Tobin's Q	55,089	2.086	1.534	0.764	1.089	1.515	2.41	8.001
CAPX ^{Control}	54,877	0.009	0.011	0	0.001	0.005	0.012	0.051
SG&A ^{Control}	55,397	0.044	0.052	0	0.004	0.027	0.063	0.22
Leverage	51,701	0.389	0.312	0	0.129	0.369	0.568	1.333
Tangibility	53,398	0.22	0.245	0.002	0.035	0.118	0.321	0.859
Profitability	55,046	0.206	0.184	0.005	0.061	0.165	0.29	0.801

Table 2: **Web Technology Adoption Determinants**

This table analyzes the correlation between web technology adoption and firm fundamentals. *Size* is the natural log of assets. *Cash Flow* is income before extraordinary items and depreciation scaled by lagged assets. *ROA* is net income scaled by lagged assets. *Tobin's Q* is market value of assets divided by book value of assets. *CAPX* is quarterly capital expenditure (derived from year-to-date measure *capxy*) scaled by lagged assets. *SG&A* is quarterly selling, general, and administrative spending (a more accurate version by Peters and Taylor (2017)) scaled by lagged assets. *Leverage* is the book leverage, defined as sum of short-term and long-term debt, divided by sum of short-term debt, long-term debt, and shareholders' equity. *Tangibility* is PP&E scaled by assets. *Profitability* is the gross profitability. The regressions include firm and quarter fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%.

	IHS(No. Tech) (1)	IHS(No. Tech ^{Analytics}) (2)	IHS(No. Tech ^{Others}) (3)
Size	0.0423** (0.0166)	0.0769*** (0.0220)	0.0408** (0.0166)
Cash Flow	0.5612** (0.2828)	0.5783 (0.4566)	0.5043* (0.2746)
ROA	-0.6656** (0.2820)	-0.6907 (0.4493)	-0.6179** (0.2745)
Tobin's Q	0.0051 (0.0053)	-0.0027 (0.0070)	0.0056 (0.0053)
CAPX	0.1277 (0.4267)	-0.6169 (0.5792)	0.1753 (0.4298)
SG&A	0.2786 (0.2795)	0.3729 (0.3577)	0.2402 (0.2799)
Leverage	-0.0479* (0.0257)	-0.0616* (0.0345)	-0.0468* (0.0257)
Tangibility	-0.0494 (0.1066)	0.0926 (0.1398)	-0.0534 (0.1062)
Profitability	-0.0514 (0.0609)	-0.1645* (0.0876)	-0.0325 (0.0608)
Observations	48,408	48,408	48,408
Adjusted R ²	0.8677	0.8295	0.8638
Firm FE	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes

Table 3: **GDPR's Impact on Data Processing Activities and Firm Fundamentals in Two-period Difference-in-differences**

This table reports two-period difference-in-differences regressions examining the impact of GDPR on firms' data processing activities and fundamentals. The regression model is

$$Dep\ Var_{i,t} = \beta_1\ High\ Legal_i \times Post_t + Firm\ FE + Time\text{-}Asset\ Quintile\ FE + \epsilon_{i,t}.$$

The dependent variable for firm i is either the inverse hyperbolic sine of the number of web technologies used in month t (Column 1), or firm characteristics in quarter $t+1$ scaled by 2016Q1 assets. Columns 2 to 6 show sales, income before extraordinary items, net income, SG&A spending, and CAPX. $High\ Legal_i$ is a time-invariant indicator equal to one if firm i 's share of legal staff is above the median in March 2016. $Post_t$ is an indicator equal to one after May 2018 (GDPR's implementation). The regressions include firm and time-by-2016Q1 asset quintile fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%.

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)	CAPX (6)
High Legal \times Post	-0.1102*** (0.0239)	-0.0118** (0.0048)	-0.0009 (0.0014)	-0.0008 (0.0014)	-0.0023* (0.0012)	0.0004 (0.0004)
Observations	179,199	53,627	53,649	53,643	53,691	53,196
Adjusted R ²	0.8358	0.8750	0.6207	0.6106	0.9052	0.6516
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes					
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes	Yes

Table 4: **Data Processing's Effect on Sales in Two-stage Least Squares**

This table reports two-stage least squares regressions examining data processing's effect on sales. Column 1 regresses sales on data processing activities using simple OLS. Column 2 is the first stage of the IV regression, regressing data processing activities on the IV, *High Legal* \times *Post*. Column 3 is the second stage of the IV regression, regressing sales on predicted data processing from the first stage. The regressions include firm and quarter-by-2016Q1 asset quintile fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%.

	Sales OLS	IHS(No. Tech) IV	Sales
	(1)	(2)	(3)
IHS(No. Tech)	0.0071** (0.0030)		0.1657** (0.0821)
High Legal \times Post		-0.0711*** (0.0221)	
Observations	53,627	53,627	53,627
Adjusted R ²	0.8750	0.8642	0.8252
F-test (1st stage)		10.3821	
Firm FE	Yes	Yes	Yes
Quarter-Asset Quintile FE	Yes	Yes	Yes

Table 5: **Heterogeneous Treatment Effects of GDPR: The Role of Workforce Composition**

This table reports heterogeneous treatment effects of GDPR. The regressions interact *High Legal* \times *Post* with indicators for ex-ante workforce composition traits. Panel **A** explores differences in treatment effects between firms with below-median share of software workforce in March 2016 (*Low SW*) and those above the median. Panel **B** explores differences in treatment effects between firms with below-median share of EU workers in March 2016 (*Low EU*) and those above the median. The regressions include firm and time-by-2016Q1 asset quintile fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%.

(A) **SW**

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)
High Legal \times Post	-0.0501* (0.0287)	-0.0020 (0.0061)	0.0010 (0.0015)	0.0011 (0.0015)	-0.0005 (0.0015)
High Legal \times Post \times Low SW	-0.1193*** (0.0346)	-0.0197*** (0.0060)	-0.0039** (0.0018)	-0.0038** (0.0019)	-0.0037** (0.0015)
Observations	179,199	53,627	53,649	53,643	53,691
Adjusted R ²	0.8364	0.8752	0.6208	0.6107	0.9053
Firm FE	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes				
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes

(B) **EU**

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)
High Legal \times Post	-0.0468 (0.0298)	-0.0048 (0.0059)	0.0010 (0.0017)	0.0010 (0.0017)	-0.0009 (0.0016)
High Legal \times Post \times Low EU	-0.1125*** (0.0346)	-0.0126** (0.0060)	-0.0035* (0.0018)	-0.0033* (0.0019)	-0.0026* (0.0015)
Observations	179,199	53,627	53,649	53,643	53,691
Adjusted R ²	0.8363	0.8751	0.6208	0.6106	0.9052
Firm FE	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes				
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes

Table 6: **Robustness Checks**

This table explores a set of robustness checks on Table 3. Panel A excludes firms headquartered in California. Panel B alternatively defines *Post* as one after GDPR's announcement in April 2016, and zero otherwise. Panel C changes the winsor levels to 5% and 95%. Panel D excludes 2020, the last year in the sample and the period of the COVID-19 crisis. The regressions include firm and time-by-2016Q1 asset quintile fixed effects, with standard errors clustered at the firm level. Continuous variables are winsorized at 2% and 98% except for Panel C.

(A) **Excluding California firms**

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)	CAPX (6)
High Legal \times Post	-0.1269*** (0.0262)	-0.0103** (0.0052)	0.0000 (0.0013)	0.0001 (0.0013)	-0.0016 (0.0012)	0.0005 (0.0004)
Observations	147,073	44,202	44,224	44,218	44,253	43,794
Adjusted R ²	0.8367	0.8816	0.6041	0.5921	0.9114	0.6613
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes					
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes	Yes

(B) **Alternative treatment cutoff date**

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)	CAPX (6)
High Legal \times Post	-0.0672*** (0.0197)	-0.0115*** (0.0043)	-0.0012 (0.0013)	-0.0012 (0.0013)	-0.0022** (0.0011)	-0.0001 (0.0003)
Observations	179,199	53,627	53,649	53,643	53,691	53,196
Adjusted R ²	0.8352	0.8750	0.6207	0.6106	0.9052	0.6516
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes					
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes	Yes

(C) **Winsor at 5% and 95%**

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)	CAPX (6)
High Legal \times Post	-0.0824*** (0.0218)	-0.0075* (0.0041)	0.0009 (0.0010)	0.0010 (0.0010)	-0.0023** (0.0010)	0.0004 (0.0003)
Observations	179,199	53,627	53,649	53,643	53,691	53,196
Adjusted R ²	0.8400	0.8848	0.6340	0.6251	0.9204	0.6846
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes					
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes	Yes

(D) Ending at 2019

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)	CAPX (6)
High Legal \times Post	-0.0900*** (0.0219)	-0.0150*** (0.0041)	-0.0013 (0.0013)	-0.0013 (0.0014)	-0.0030*** (0.0011)	0.0002 (0.0003)
Observations	153,871	47,088	47,105	47,100	47,141	46,709
Adjusted R ²	0.8517	0.8950	0.6388	0.6293	0.9168	0.6809
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes					
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes	Yes

A Additional Results

Table A.1: **Treatment Allocation Determinants**

This table analyzes the correlation between *High Legal* and firm fundamentals in 2016Q1 using firm-level regressions. *High Legal* is an indicator equal to one if a firm's share of legal worders is above median in March 2016. *Size* is the natural log of assets. *Cash Flow* is income before extraordinary items and depreciation scaled by lagged assets. *ROA* is net income scaled by lagged assets. *Tobin's Q* is market value of assets divided by book value of assets. *CAPX* is quarterly capital expenditure (derived from year-to-date measure *capxy*) scaled by lagged assets. *SG&A* is quarterly selling, general, and administrative spending (a more accurate version by [Peters and Taylor \(2017\)](#)) scaled by lagged assets. *Leverage* is the book leverage, defined as sum of short-term and long-term debt, divided by sum of short-term debt, long-term debt, and shareholders' equity. *Tangibility* is PP&E scaled by assets. *Profitability* is the gross profitability. Column 2 includes NAICS 2-digit fixed effects. All standard errors are clustered at the firm level and continuous variables are winsorized at 2% and 98%.

	High Legal	
	(1)	(2)
Constant	0.4191*** (0.0542)	
Size	0.0349*** (0.0062)	0.0257*** (0.0063)
Cash Flow	-2.5640** (1.3040)	-0.8032 (1.2822)
ROA	1.4640 (1.3094)	-0.1586 (1.2925)
Tobin's Q	0.0148 (0.0097)	0.0276*** (0.0100)
CAPX	-0.7824 (1.2444)	-0.1493 (1.2203)
SG&A	-1.2418*** (0.2973)	-0.7957** (0.3122)
Leverage	-0.0054 (0.0324)	-0.0230 (0.0318)
Tangibility	-0.0116 (0.0587)	0.0425 (0.0775)
Profitability	-0.6025*** (0.0809)	-0.2651*** (0.0886)
Observations	1,868	1,866
Adjusted R ²	0.1557	0.2707
NAICS2 FE		Yes

Table A.2: **GDPR's Impact on Data Processing Activities and Firm Fundamentals in Two-period Difference-in-differences, with Control Variables**

This table reports two-period difference-in-differences regressions examining the impact of GDPR on firms' data processing activities and fundamentals. The regression model is

$$Dep\ Var_{i,t} = \beta_1\ High\ Legal_i \times Post_t + Firm\ FE + Time-Asset\ Quintile\ FE + Controls_{i,t} + \epsilon_{i,t}.$$

The dependent variable for firm i is either the inverse hyperbolic sine of the number of web technologies used in month t (Column 1), or firm characteristics in quarter $t + 1$ scaled by 2016Q1 assets. $High\ Legal_i$ is a time-invariant indicator equal to one if firm i 's share of legal staff is above the median in March 2016. $Post_t$ is an indicator equal to one after May 2018 (GDPR's implementation). $Controls_{i,t}$ are defined below. $Size$ is the natural log of assets. $Cash\ Flow$ is income before extraordinary items and depreciation scaled by lagged assets. ROA is net income scaled by lagged assets. $Tobin's\ Q$ is market value of assets divided by book value of assets. $CAPX$ is quarterly capital expenditure (derived from year-to-date measure $capxy$) scaled by lagged assets. $SG\&A$ is quarterly selling, general, and administrative spending (a more accurate version by [Peters and Taylor \(2017\)](#)) scaled by lagged assets. $Leverage$ is the book leverage, defined as sum of short-term and long-term debt, divided by sum of short-term debt, long-term debt, and shareholders' equity. $Tangibility$ is PP&E scaled by assets. $Profitability$ is the gross profitability. The regressions include firm and time-by-2016Q1 asset quintile fixed effects, with standard errors clustered at the firm level and continuous variables winsorized at 2% and 98%.

	IHS(No. Tech) (1)	Sales (2)	IB (3)	NI (4)	SG&A (5)	CAPX (6)
High Legal \times Post	-0.0710*** (0.0229)	-0.0147*** (0.0036)	-0.0014 (0.0013)	-0.0010 (0.0013)	-0.0029*** (0.0010)	0.0000 (0.0002)
Size	0.0514*** (0.0166)	0.1492*** (0.0058)	-0.0145*** (0.0018)	-0.0150*** (0.0018)	0.0345*** (0.0016)	0.0076*** (0.0003)
Cash Flow	0.3945 (0.2913)	0.1508* (0.0814)	0.1236*** (0.0393)	0.0761* (0.0462)	0.0701** (0.0325)	-0.0113* (0.0058)
ROA	-0.4820* (0.2910)	-0.1434* (0.0803)	0.0788** (0.0385)	0.1292*** (0.0461)	-0.0693** (0.0332)	0.0196*** (0.0055)
Tobin's Q	0.0038 (0.0053)	0.0129*** (0.0014)	0.0049*** (0.0006)	0.0051*** (0.0006)	0.0015*** (0.0004)	0.0010*** (0.0001)
CAPX	0.1018 (0.4232)	0.3222*** (0.1103)	0.1345*** (0.0431)	0.1282*** (0.0449)	0.0200 (0.0217)	0.4792*** (0.0165)
SG&A	0.2358 (0.2737)	-0.1604** (0.0752)	-0.0920*** (0.0269)	-0.0852*** (0.0277)	0.3155*** (0.0352)	-0.0037 (0.0051)
Leverage	-0.0320 (0.0261)	0.0215*** (0.0081)	0.0006 (0.0026)	0.0013 (0.0026)	0.0011 (0.0019)	-0.0019*** (0.0005)
Tangibility	-0.0060 (0.1062)	-0.0108 (0.0254)	-0.0228*** (0.0074)	-0.0273*** (0.0075)	-0.0066 (0.0078)	-0.0023 (0.0021)
Profitability	-0.0365 (0.0614)	0.2762*** (0.0268)	0.0269*** (0.0075)	0.0247*** (0.0076)	-0.0098* (0.0053)	-0.0029*** (0.0011)
Observations	47,359	47,156	47,160	47,158	47,184	47,135
Adjusted R ²	0.8688	0.9073	0.6455	0.6359	0.9308	0.7243
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month-Asset Quintile FE	Yes					
Quarter-Asset Quintile FE		Yes	Yes	Yes	Yes	Yes