Working Paper 21-122

Data Governance, Interoperability and Standardization: Organizational Adaptation to Privacy Regulation

Sam (Ruiqing) Cao
Marco Iansiti

Harvard
Business
School

Data Governance, Interoperability and Standardization: Organizational Adaptation to Privacy Regulation

Sam (Ruiqing) Cao
Stockholm School of Economics

Marco Iansiti
Harvard Business School

Working Paper 21-122

We thank Disha Monga and Jiayu Lih for excellent research assistance. We are grateful to Alexandria Sheng, Ellen Wu, Amanda Pratt, Ross Sullivan, and Keystone Strategy for implementing the survey and collecting the data. Satya Nadella, Kinney Zalesne, and Greg Shaw at Microsoft provided invaluable input and helped fund the data collection process. Marco Iansiti has worked with many companies in the IT sector and in traditional industries on the topic of digital transformation.

Funding for this research was provided in part by Harvard Business School.

Data Governance, Interoperability and Standardization: Organizational Adaptation to Privacy Regulation

Sam (Ruiqing) Cao[†]
Stockholm School of Economics
sam.cao@hhs.se

Marco Iansiti
Harvard Business School
miansiti@hbs.edu

Abstract

The increasing availability of data can afford dynamic competitive advantages among data-intensive corporations, but governance bottlenecks hinder data-driven value creation and increase regulatory risks. We analyze the role of two technological features of data architecture that facilitate internal data governance – Application Programmatic Interfaces (APIs) that publish interdepartmental data and standardization of identity and access management (IAM) software – in shaping large data-intensive corporations' adaptation to privacy regulation. Using annual establishment data for the largest U.S. financial services corporations and the enforcement of the General Data Protection Regulation (GDPR) in 2018 as a natural experiment, we show that internal data APIs and standardization of IAM software significantly mitigate establishments' revenue loss and IT budget reduction in response to GDPR enforcement. Compliance costs measured by IT hiring increased substantially after GDPR enforcement only for firms without internal data APIs. Our findings highlight the importance of interoperability and standardization as technical conditions that facilitate dynamic integrative capability, allowing large data-intensive corporations to ensure proper data governance and adapt to privacy regulation.

[†] Corresponding author. Email: sam.cao@hhs.se. We are grateful to Alexandria Sheng, Ellen Wu, Amanda Pratt, and Ross Sullivan of Keystone Strategy for implementing the survey and collecting the data, and to Satya Nadella, Kinney Zalesne, and Greg Shaw at Microsoft for providing invaluable input and funding data collection. We are grateful to many individuals and especially Tobias Kretschmer, Magnus Mähring, Kalle Lyytinen, and Robert Wayne Gregory for constructive feedback that helped shape the paper, and to participants and especially discussants at the MISQ Author Development Workshop, ZEW Conference on the Economics of ICT, Organization Design Community Idea Development Workshop, CESifo Area Conference on Economics of Digitization, Digital Economy Workshop, Munich Summer Institute, Strategy Science Conference, LMU ISTO Seminar, and the IRIS/SCIS Conference for very helpful comments. All errors are our own.

1. Introduction

The pervasive presence of digital technologies increasingly disrupts traditional industries, leading to fiercer market competition and heightened consumer expectations (Vial, 2019; Kretschmer & Khashabi, 2020; Drechsler, Gregory, Wagner, & Tumbas, 2020). As a result, many legacy corporations embarked on digital transformation, aiming to use data to improve internal process efficiency, deliver higher-quality products and services, and develop innovative ML/AI solutions to automate decision-making and address customer needs (Loebbecke & Picot, 2015). At the same time, firms' aggressive exploitation of personal data raises ethical and societal concerns about data privacy and information security. Privacy regulations such as the European Union's General Data Protection Regulation (GDPR) led to global repercussions for firms that handle personal data (Peukert, Bechtold, Batikas, & Kretschmer, 2022; Johnson, 2022). These industry and macro-level trends shape an increasingly turbulent external environment in which today's large corporations find themselves, and their ability to adapt depends on proper internal data governance.

Integrative capability (Helfat & Raubitschek, 2016; Helfat & Raubitschek, 2018; Vial, 2019) is a crucial bottleneck for internal data governance, defined by the ability to incorporate changes to its products, resources, capabilities, and business models reliably and efficiently. Integrative capability facilitates linkages between siloed subsystems and enables intraorganizational communication and coordination. When organizations face regulatory shocks requiring stricter system-wide data governance, integrative capability can ensure efficient adaptation at a low cost (e.g., Agrawal, Gans, & Goldfarb, 2023). However, integrative capability does not simply result from strategic choices that organizations make from the top down. Instead, it requires technological features within the enterprise architecture that may be highly inertial. To date, relatively little is understood about what technological factors contribute to integrative capability, and most studies hinting at these factors have been theoretical (Drechsler, Gregory, Wagner, & Tumbas, 2020).

We aim to fill this research gap by investigating technological features contributing to integrative capability and facilitating internal data governance. We ask two research questions: Which technological features can facilitate integrative capability within large data-intensive corporations? How do these technological features affect large data-intensive corporations' ability to adapt to privacy regulations that mandate stricter internal data governance, e.g., through mitigating negative repercussions for business performance and reducing compliance costs?

To investigate these questions, we examine an empirical setting around the General Data Protection Regulation (GDPR) enforcement in 2018, which mandated stricter internal data governance requiring visibility into an organization's internal data sources and incorporation of compliance solutions with existing technologies. We propose two channels for effective organizational adaptation to privacy regulation. *First*, internal data interoperability enables organizations to meet regulatory mandates for data tracking and third-party audits. Organizations can orchestrate data interoperability using Application Programming Interfaces (APIs) to publish data across departments and functional units. Internal data APIs specify the technical and governance rules for organizational members to access the same underlying data without friction, automating the exchange of information across different functional units. They enable the organization to combine information across different data sources and thus gain visibility of the entire data system.

Second, standardization of identity and access management (IAM) software components ensures low-cost and efficient adaptation to privacy regulation by facilitating scalable compliance solutions that can be applied globally. When software components require updates to incorporate compliance solutions or new features to ensure compliance, standardization allows the compliance solution developed for a standardized technology component to be easily re-used by other functional units or subsystems. These technological features – internal data APIs and standardization of IAM software – contribute to integrative capability and enable organizations to adapt effectively to privacy regulation. While interoperability enables internal data tracking, standardization allows efficient global scaling of compliance solutions. Without these features, organizations may reduce value-enhancing IT investments to lower compliance risks and compromise on revenue losses to comply with the regulation. They may also incur higher compliance costs by hiring IT workers to integrate data manually and perform regulation-related tasks.

We test these predictions by estimating the causal impact of GDPR enforcement on organizational performance and compliance costs. The enforcement of the GDPR in 2018 provides a natural experiment that allows us to identify mechanisms of organizational adaptation to a regulatory change that mandates stricter internal data governance. We use a triple differences regression framework and conduct extensive robustness checks, including matching methods such as PSM (propensity score matching) and CEM (coarsened exact matching), subsample difference-in-differences analyses, and synthetic control DID methods on aggregate corporation-level data. Our empirical sample consists of annual establishment observations from twenty-five of the largest U.S. corporations with an average founding year of 1905, which accounted for more than 36% of the total gross output of the entire U.S. finance and insurance sector. We combine establishment-level data on revenue, IT investments, and software products from Aberdeen CI Technology Database (CITDB) with three supplemental data sources: Burning Glass Technologies (BGT) data on job postings, Keystone-Microsoft survey on data architecture, and the corporations' public annual reports (10-K forms) downloaded from the U.S. Securities and Exchange Commission (SEC) website.

The final data set contains annual observations of a balanced panel of 17,311 establishments from 2016 to 2020. We identify each sample corporation's exposure to GDPR by using information from annual reports to derive the extent to which their revenues are exposed to the European market and require handling personal data. We find robust empirical evidence that the availability of APIs for publishing internal data and the standardization of identity and access management (IAM) software mitigate establishments' performance decline due to GDPR enforcement. We also find evidence that these technological features lower IT investments following GDPR enforcement. Furthermore, organizations without internal data APIs significantly increase compliance costs by hiring more IT workers, while other firms do not incur higher labor costs in response to GDPR enforcement.

Our findings provide important insights into the factors contributing to and implications of internal data governance for an organization designing a digital strategy to create and capture value from data-driven innovation. Interoperability and standardization are crucial to facilitating integrative capability within large organizations. They increase organizations' capacity to respond to system-wide regulatory shocks that require stricter internal data governance. When these technological conditions are not met, organizations may incur higher labor costs to comply with the regulation and scale back value-enhancing IT investments that risk compliance violation because they cannot orchestrate automated and error-robust mechanisms for data integration that meet regulatory requirements. Hence, our results point to these technical aspects of the enterprise data architecture as critical bottlenecks that may hinder value creation and capture from data-driven innovation such as analytics and ML technologies at the implementation stage. Our results also add to the understanding of heterogeneity in the impact of privacy regulation on large data-intensive corporations. We show that architectural and technological capabilities contributing to internal data governance can explain the variation in organizational adaptation to regulatory risks.

2. Theory Development

Digital technologies have brought disruptive changes to traditional industries and altered the competitive landscape for many large incumbent corporations. New business models increasingly emerge from the availability of data and digital technologies, replacing traditional pathways of value creation and capture (Vial, 2019; Piccoli, Rodriguez, & Grover, 2023). When customers interact with products and services through digital technologies such as IoT devices and social media (Barrett, Davidson, Prabhu, & Vargo, 2015), the process generates large amounts of data that allow firms to deliver efficient and high-quality services to customers. For legacy corporations in traditional industries, becoming “customer-centric” is one of the primary motivators for digital transformation (Kolbjornsen & Rockwood, 2019; Elm, Gaughan, & Brown, 2021). Companies that successfully develop solutions based on large-scale user data can enhance their business performance with predictive analytics and AI capabilities (e.g., Wu, Hitt, & Lou, 2020;

Bessen, Impink, Reichensperger, & Seamans, 2020; Gregory, Henfridsson, Kaganer, & Kyriakou, 2021; Berman & Israeli, 2022).

Data is an important source of dynamic capability that enables firms to assemble resources and leverage emergent technologies swiftly in response to environmental changes (Teece, 2007; Helfat & Raubitschek, 2018; Vial, 2019). However, large legacy corporations often fail to create value from their data assets. Their siloed and differentiated data systems constitute bottlenecks for modularizing data as digital capabilities that can be shared with other organizational members and functional teams. For these organizations, internal data governance is essential for ensuring the proper delivery of data as modularized resources. Data governance requires intraorganizational coordination and integrative capability, defined as "reliable, repeatable communication and coordination activity directed toward the introduction and modification of products, resources, capabilities, and business models… and encompass the capacity to establish and alter how communication and coordination activities take place" (Helfat & Raubitschek, 2018; Vial, 2019).

For large legacy corporations, internal data governance requires technical capabilities associated with the enterprise architecture (Mithas, Tafti, & Mitchell, 2013; Bharadwaj, El Sawy, Pavlou, & Venkatraman, 2013). Corporations have recognized the importance of a solid architectural foundation for internal data governance and ensuring data quality and representation, as the following quote by the CIO at one of the largest U.S. banks illustrates:

> *"You can hire a consulting company that will tell you, 'Oh, there's a big opportunity in externalizing services.' You cannot wake up one day and start externalizing stuff because you can only do it if you have a solid foundation underneath… We've invested a lot in our data quality, in our data lineage, in our data platforms, the ability to abstract the complexity of data, how we represent complex option, and how we represent trades. We decided to take a fairly modern and standardized approach to that by creating a standard which started internally." – Partner and CIO at Goldman Sachs*

Technological architecture can be a critical bottleneck for large legacy corporations, especially as they move into the digital era, where new digital-enabled business models reshape these organizations with increasingly decentralized infrastructural technologies and distributed data systems (Henderson & Clark, 1990; Tilson, Lyytinen, & Sørensen, 2010; Albert & Siggelkow, 2022). As large corporations embark on transformation programs (Kretschmer & Khashabi, 2020; Wessel et al., 2021), they may need to pursue existing and new business models simultaneously to maintain business continuity. Existing firm assets can become barriers to transformation when the adjustment costs associated with sharing resources across existing and new business models are particularly high (Eklund & Kapoor, 2019). Corporations must replace existing architectural knowledge to establish a new architectural foundation, which can encounter tremendous resistance and organizational inertia.

Internal APIs that publish interdepartmental data are a crucial element of the architectural foundation that can facilitate flows and linkages across disparate subsystems and data sources. These data APIs are automated interfaces that ensure low-cost and robust data sharing across different parts of the organization, thus ensuring appropriate modularization of scalable data sources across distributed systems that increasingly characterize digitalized organizations. For example, UnitedHealth used application programmatic interfaces (APIs) to enable the re-use of the same underlying data across different applications (Optum, 2017). JP Morgan Chase used APIs to facilitate instant payment across multiple locations (J.P. Morgan, 2016). MoneyGram overhauled its IT infrastructure using APIs to streamline operational processes and improve the quality of customer experiences (Business Reporter, 2021). The following quotes from a Bloomberg news article and a technology executive at one of the largest U.S. multinational insurance corporations illustrate the importance of APIs for breaking data siloes and ensuring consistency in data sources and applications across the entire IT infrastructure at a low cost.

> *"Forward-thinking CIOs are freeing their data from isolated back-end systems to next-generation platforms, much like successful retail organizations… Application programming interfaces (APIs) enable many applications to consume data. It's a very modern, very powerful and ultimately a cost-savings approach to reuse data and create consistency of data across all apps as data is externalized." – CIO of UnitedHealth Group*

> *By June 2020, MoneyGram had already built a direct-to-consumer digital channel that provides an immersive experience that rivals those of many leading e-commerce brands. The company also proactively overhauled its supporting IT infrastructure, modernized its APIs and streamlined its operating model to support the growth of digital. – Bloomberg Business Reporter (MoneyGram)*

Coordination and centralization are crucial to internal data governance, opposite to the prevailing logic of decentralized digital organizations. The modular design principle (Sanchez & Mahoney, 1996; Baldwin

& Clark, 2000; Ethiraj & Levinthal, 2004) underpins the digital organization to facilitate the benefits of loosely coupled components through flexible innovation and low-cost scaling (Simon, 1962; Langlois & Robertson, 1992; Ulrich, 1995; Campagnolo & Camuffo, 2010; Zakerinia & Yang, 2023). Management of modularized organizations is often naturally decentralized as environmental changes can be addressed by confining the adaptation to localized solutions that do not involve other parts of the organization (e.g., Englmaier, Galdon-Sanchez, Gil, & Kaiser, 2019; Aghion et al., 2021). However, when the external changes are regulatory or governance-oriented, they require the entire organization to respond in a coordinated fashion.

Standardization of technology components can reduce the friction of scaling a particular software component or sharing a data API globally across a large organization. Standardization requires centralized decisions about which technology components are used and where to source them. Once implemented, standardization lowers the structural and cognitive complexity of the enterprise architecture (Xia & Lee, 2005; Widjaja & Gregory, 2012) and, thus, the costs of coordination and communication across subsystems. In a dynamic environment, organizations must go beyond local responses to regulatory adaptation, as intra-organizational spillovers of governance bottlenecks require organization-wide coordination to address (e.g., Agrawal, Gans, & Goldfarb, 2023). The following quotes from public interviews with two technology executives at Fortune 500 corporations illustrate that standardization from a centralized perspective reduces scaling frictions and can particularly help organizations adapt to regulatory requirements by enabling the efficient scaling and re-use of compliance solutions globally.

> *"We've had to strike a balance between what tools are regional and what is controlled centrally. Whatever we can standardize globally reduces friction and helps us get to market even faster." – Executive VP & CIO of Prudential Financial*

> *"One of the primary advantages of doing this from a global perspective is it gives you a considerable opportunity to leverage scale. What we generally find is that when we build a global solution, generally between 70% and 80% of that solution ends up being reusable. There are some things like regulatory requirements and privacy laws that are specific to a particular geography, but we usually have the opportunity to scale a majority of our solution with considerable speed." — Executive VP of Global Technology and Operations at Metropolitan Life Insurance Company (MetLife)*

2.1 GDPR Enforcement, Data Governance, and Business Performance

Digital firms' aggressive exploitation of personal data raises concerns about privacy and information security. The potential of generative innovation from the hyper-scaling of data amplifies the downside of the potential abuse of personal information. Stakeholders, including policymakers and consumers, have become increasingly aware of the societal risks associated with firms' collection and usage of personal data. Among the most profound regulatory changes in recent years is the General Data Protection Regulation (GDPR), introduced in 2016 and enforced in May 2018 by the European Union. Since then, different regulatory frameworks have emerged worldwide, including the California Consumer Privacy Act (CCPA) and other regional laws that implement GDPR-style frameworks for protecting personal data rights.

The GDPR frames data privacy protection around mandating consumer consent for data sharing, hence putting the control of personal data in individuals' own hands (Johnson, 2022). The GDPR profoundly impacted organizations within and outside Europe (Peukert, Bechtold, Batikas, & Kretschmer, 2022), and can particularly affect global firms that operate outside Europe but sell products and services to European consumers. The GDPR drastically expanded the scope of privacy regulation and increased the financial penalties associated with violations relative to existing laws. The announcement of the GDPR subjected firms to substantial uncertainty around whether their current data practices were compliant with the regulation. It exposed organizations to compliance risks as they built new technology systems to deliver customer-focused services and data-driven innovations.

The GDPR may have especially salient effects on large multinational corporations in data-intensive sectors. Large corporations are likelier to commit to enforcement efforts than small and medium-sized enterprises (SMEs) and startups because they can potentially be fined a hefty amount due to their massive global revenues. Large firms are also more likely to have historically installed processes and resources for adapting to regulatory changes. Large corporations are also under intensive legitimacy scrutiny from regulators and consumers relative to SMEs and startups. Recent research suggests that public U.S. firms increased their attention to data privacy due to exposure to the GDPR (Boroomand, Leiponen, & Vasudeva, 2022). In our data, corporations with a larger share of their businesses involving European customers discussed GDPR in their annual reports (10-K forms) in greater detail since the regulation was introduced in 2016 (Appendix Figure B2). Meanwhile, the largest U.S. financial services corporations have recorded zero fines due specifically to GDPR violations up to 2020, according to a crowdsourced database on companies fined for GDPR violations.[1]

At least two major problems affect large multinational corporations in complying with the GDPR. First, large corporations own many geographically scattered establishments, heterogeneous product lines, and diverse customer segments. They may serve customers in multiple locations and use an identical copy of personal data co-existing across jurisdictions where different privacy laws apply. The transfer of personal information from Europe to other countries is governed by the Standard Contractual Clauses (also known as the EU Model Clauses). For data to be allowed to flow across national borders for processing (Article 44 of the GDPR), both corporations (as data controllers) and infrastructure providers (as data processors) must meet the conditions of the GDPR (Article 44). Second, under the GDPR, consumers as data subjects have the right to obtain copies, request changes and deletions, and restrict the processing and use of their data. In line with this objective, firms must conduct data audits and track data sources and data flows across their internal systems.

Corporations must adapt to the changing regulatory environment and uncertainty regarding externally mandated data governance standards. Otherwise, failing to comply with the GDPR can risk their legitimacy in the eyes of stakeholder audiences, who may withhold resources from these firms. Users may stop paying for services if they believe the firm misused their data. Regulators may issue hefty fines if they deem the firms' data practice to be at odds with the regulation. Failure to develop compliance solutions that can be applied widely across the system can force corporations to remove technology components that demonstrate non-compliant data storage and processing practices, lowering service quality and customer experience. As

[1] GDPR Enforcement Tracker: https://www.enforcementtracker.com

a result, less effective data-driven capabilities and loss of digital innovation decrease consumer demand and lower revenues.

2.1.1 Effect of Application Programmatic Interfaces (APIs) and Data Interoperability

Data silos are an urgent organizational problem that plagues many legacy incumbent corporations. Distributed systems and infrastructure can lead to increasing differentiation in the enterprise architecture, causing barriers to communication and coordination around joint tasks across disparate subsystems. This problem is exacerbated by forces of digitalization that pressures traditional corporations to transition from a hierarchical and centralized form of organization into a distributed and decentralized system (Eklund & Kapoor, 2019; Giustiziero, Kretschmer, Somaya, & Wu, 2022).

Internal APIs can break down data silos by publishing data across departments by specifying technical and governance properties that clearly define how the functionalities encapsulated within the APIs can be shared with any organizational member with appropriate access rights. The availability of such internal data APIs facilitates integrative capability (Helfat & Raubitschek, 2018; Vial, 2019), thus combining distributed resources across different parts of the organization. APIs expose data as digital capabilities and allow users in other functional units to access these capabilities (Melville & Kohli, 2021; Benzell, Hersh, & Van Alstyne, 2022; Piccoli, Rodriguez, & Grover, 2023). Therefore, they enable multiple teams to jointly work on the same underlying data across different subsystems without introducing technical complexities and errors that can hinder data quality.

Internal data APIs allow organizations to satisfy the GDPR mandates for internal data governance that require visibility into the entire data system for tracking information flows and conducting third-party audits. Data silos hinder regulatory compliance by preventing information flows across the organization and interdepartmental coordination for verifying underlying data sources used by multiple teams. APIs automate the exchange of data and information across subsystems, therefore breaking down data silos and linking data across subsystems much more efficiently than labor-intensive manual processes. On the other hand, organizations that do not have internal data APIs cannot easily comply with the GDPR. Thus, they may reduce investments in value-enhancing digital technologies to lower the risks of violating the regulation. Such adaptation may constrain organizational performance and lower business revenues. Therefore, we propose the following hypotheses.

Hypothesis #1a: Internal data APIs improve business performance following GDPR enforcement which imposes stricter data governance across a large corporation's entire internal data system.

Hypothesis #1b: Data-intensive large organizations without internal data APIs increase IT labor costs following GDPR enforcement to perform compliance-related tasks.

Hypothesis #1c: Data-intensive large organizations without internal data APIs reduce investments in IT assets following GDPR enforcement.

2.1.2 Effects of Vendor Standardization and Interoperability of Compliance Solutions

In complex architectures that characterize large legacy corporations' data systems, high degrees of differentiation among existing technology components raise IT costs and introduce frictions that hinder adaptation (Xia & Lee, 2005; Widjaja & Gregory, 2012). Heterogeneity in the enterprise architecture results from many historical events (e.g., mergers and acquisitions) that reflected past organizational decisions but no longer fit present-day needs. These differentiated systems make it difficult for organizations and functional sub-units to set up automated processes that respond swiftly to changing regulatory mandates because subsystems do not share the same technology components and communication standards.

On the other hand, standardization of technology components tame enterprise architecture complexity and lower IT costs (Boh & Yellin, 2006). Standardization facilitates technological conditions conducive to interoperability. When data and software applications use similar vendors and standardized specifications across different subsystems, it lowers the barriers to making coordinated organization-wide changes to the system and incorporating new technologies that can be easily scaled across the organization. Technology executives at large financial services corporations have recognized the importance of standardization to improve dynamic integrative capability, which allows them to introduce compliance solutions and scale them across the entire organization. For example, Morgan Stanley and Goldman Sachs manage their infrastructure by creating shared internal standards and extracting standardized assets from an integrated operator (Infrastructure Investor, 2021; High, 2023). Prudential Financial standardizes new software solutions and rolls them out efficiently at a global scale (Noyes, 2021).

Identity and access management (IAM) software products are essential to data governance as they grant users at different levels appropriate access to different internal systems and data sources. Standardizing IAM software can lower the costs of complying with the GDPR because organizations can re-use a compliance solution developed for a particular technology component (e.g., a standardized IAM software product) by directly applying it to other subsystems that use the same technology component. Standardization simplifies the adaptation and mitigates the challenges associated with highly differentiated subsystems that may require complex customized compliance solutions. For example, MetLife developed a global compliance solution for GDPR and rolled it out across the entire organization at a low cost using a standardized approach that leverages scale (Lippert & Kane, 2017). More generally, technology components produced by a vendor with a higher industry adoption rate are more standardized, and regulatory compliance solutions for these components are easier to develop and more likely to be available through the vendor.

Organizations with low standardization may struggle to deliver compliance solutions that cover highly differentiated subsystems and meet regulatory standards across all the customized solutions. Organizations failing to develop solutions that satisfy GDPR requirements before the enforcement deadline may remove non-compliant IT services and products that interact with and create value from consumer data. As a result, GDPR enforcement can lower business revenues for these organizations. Therefore, we propose the following hypotheses.

Hypothesis #2a: Standardization of identity and access management (IAM) software improve business performance following GDPR enforcement which imposes stricter data governance across a large corporation's entire internal data system.

Hypothesis #2b: Data-intensive large organizations with low standardization of IAM software reduce investments in IT assets following GDPR enforcement.

3 Data and Measures

Our data consists of annual observations of all the establishments in the Aberdeen CI Technology Database (CITDB) with available data between 2016 and 2020 that belong to some of the largest U.S. financial services corporations. In 2017, the sample corporations accounted for 36.2% of the entire U.S. Finance and Insurance (NAICS code 52) sector's total gross output.[2] The full list of these corporations is available in Table 1. These corporations engage in at least one of the four data-intensive business segments: payment, consumer banking, insurance, and asset management. Some corporations operate in multiple segments – for example, Bank of America has both a consumer banking segment and an asset management segment. Individual customers provide their personal information with varying degrees of sensitivity in exchange for products and services.[3] All the corporations were founded before the Internet era (e.g., before Amazon was founded in 1994), and the median founding year was 1905. They tend to have tremendous inertia and old systems that date back to long before digital business models came into existence. For example, the sample includes the four largest U.S. banks – Bank of America, JPMorgan Chase, Citigroup, and Wells Fargo. The final data consists of establishments of 24 corporations – all except Mastercard, which has no data in 2020 and hence all observations must be dropped to ensure the balanced panel structure of the final data.

Our outcome variable is firm performance measured by annual establishment revenue, using data from the Aberdeen CI Technology Database (CITDB). This database is the most authoritative data source on IT products and investments at the establishment level, and has been widely used across many previous studies

[2] This statistic is derived from the U.S. Bureau of Economic Analysis data on Annual GDP by Industry in 2017 (see https://apps.bea.gov/histdata/fileStructDisplay.cfm?HMI=8&DY=2017&DQ=Q4&DV=Quarter&dNRD=April-19-2018). Gross output is the appropriate denominator for calculating the share of combined revenues of the sample corporations instead of value-added which leads to a larger estimate of 61% (see https://www.bea.gov/help/faq/1197).

[3] While an individual may only need to provide information about their identities such as social security number to open a bank account, purchasing insurance may require more sensitive personal information such as medical records.

published across information systems, management, and finance journals (e.g., Forman, Goldfarb, & Greenstein, 2005; Chwelos, Ramirez, Kraemer, & Melville, 2010; Bloom, Sadun, & Van Reenen, 2012; Kretschmer, Miravete, & Pernías, 2012; McElheran, 2014; Nagle, 2019; Tuzel & Zhang, 2021; Huang, Ceccagnoli, Forman, & Wu, 2022; Sambhara, Rai, & Xu, 2022). We focus on the CITDB data that covers all U.S. establishments between 2016 and 2020. There have been documented data quality challenges with earlier years of the CITDB data (Levy, 2015, 2019), primarily due to changes in data ownership around 2016, when Aberdeen split from Harte Hanks was previously the owner company of the data provider since 2006. After substantial data checks (see Appendix Section A3 for a detailed account of data construction methodology and measurement issues), we believe the data is consistently measured and variables similarly defined during the sample period (2016 – 2020).

Our empirical analyses also investigate compliance costs and firms' actions associated with adjusting to the regulation. For this purpose, we examine two sets of outcomes: IT investments which measures actions associated with removing value-enhancing digital technologies, and IT hiring which measures labor costs associated with compliance efforts. To measure IT investments, we use annual establishment budgets for IT services, hardware, and software in the CITDB. To measure IT hiring, we use quarterly number of job postings in the Burning Glass Technologies (BGT) at the firm-state level. This includes jobs in all computer occupations except support specialists (2010 SOC codes 15-11XX except 15-115X). We also measure the number of job postings restricted to computer occupations likely to involve data (2010 SOC codes 15-111X research scientists and 15-112X analysts), and computer occupations likely to perform regulation-related tasks (Trebbi, Zhang, & Simkovic, 2023).

We supplement these key outcomes with two corporation-level data sources. One of them is a survey developed by Keystone Strategy and Microsoft jointly that contains information about the availability of application programmatic interfaces (APIs) at the corporation level. The other supplemental data source is corporations' mandatory annual reports (10-K forms) where they disclose information about market exposure by geography. For more detailed descriptions of all the data sources and how we identify establishments and combine data at different levels (from corporation to its sub-units), see Appendix Sections A1 and A2.

3.1. Measuring the Availability of Application Programmatic Interfaces (APIs)

In 2020, Keystone Strategy LLC and Microsoft Corporation jointly designed and administered a survey about firms' AI technology intensity and data architecture maturity, based on many years of deep industry expert knowledge in the space of digital transformation. The survey was conducted for non-academic purposes unrelated to this paper. The survey team reached out to senior-level technology executives (e.g., CIOs, CDOs) at legacy corporations, who oversee and manage the entire organization's IT systems and digital infrastructure. The survey question covers whether the organization permits data flows across subunits, which is typically made technical feasible through internal usage of application programmatic interfaces (APIs).

The availability of application programmatic interfaces (APIs) is measured using the following survey question which elicits responses at the corporation level: *"Do you use services interfaces or APIs to publish departmental (e.g., not application) information between departments and/or functional areas?"*. The variable is coded as *1* if the recorded response is *Yes* to this question, and *0* otherwise. Appendix Table C1 lists all the corporations by the availability of APIs, defined by the binary responses to the survey question. 7 out of 25 corporations (28%) responded *No* and hence are assigned values of *0* on the API variable.

3.2. Measuring Vendor Standardization of Technologies Subject to Privacy Regulations

We measure technology component standardization at the vendor level specifically for the identity and access management (IAM) software product category, which is essential to internal data control as it governs which users have access to which data in the organization's critical data systems. We leverage CITDB data to measure the share of all establishments that use a particular product vendor as their IAM software provider across the entire financial services sector ($IndustryShare_j$). We then calculate standardization as the average industry share across all IAM vendors present at the establishment. Equation

1 defines the establishment-level standardization variable, which is calculated using data in 2017, prior to GDPR enforcement.

$$Standardization_i = \frac{1}{|V(i)|} \sum_{j \in V(i)} IndustryShare_j \tag{1}$$

In Equation 1, each IAM software vendor is denoted by j, and all vendors present at establishment i in 2017 is denoted by the set $V(i)$. Note that each vendor can have multiple products in the IAM software category, which are often complementary solutions that work together. Therefore, $IndustryShare_j$ is given by the maximum of industry adoption rates across all products by the same vendor j in the IAM software category. An establishment with a larger standardization measure chooses more popular or established vendors that are widely used by other establishments within the financial services sector. Furthermore, when we conduct robustness checks and other analyses involving aggregate data, we calculate the average of the standardization measure across all establishments at the analysis unit level.

3.3. Measuring Corporations' Exposure to the European Consumer Market

We conduct an event study around the enforcement of GDPR in 2018 to estimate the treatment effect of GDPR enforcement on establishment revenue. Multinational corporations in the financial services sector are substantially exposed to compliance risks due to the GDPR if they operate in the European consumer market. For example, insurance companies and asset management firms serve individual clients who travel across different geographic locations including European countries. Cross-border transfer of funds (e.g., for credit card and other payment businesses, as well as consumer banking) are regulated by the GDPR and hence should incorporate enhanced privacy and security features to ensure compliance.

Corporations' operations outside Europe are affected by the GDPR due to the global nature of most business segments in the financial services sector, and the impact is larger for those that are more involved and generate a higher fraction of revenues in the European market. More generally, the treatment variable is defined by the extent to which the revenue performances of the corporations' U.S. branches are affected by regulatory risks imposed by GDPR enforcement. We measure each corporation's exposure to GDPR compliance using the first available source of information among the following (and in the order listed): (1) share of revenues generated by consumer business segments operating in Europe[4], (2) total credit risk exposure to European countries as a fraction of total interest earning assets, (3) insurance company's share of net premiums earned in Europe, and (4) share of revenues generated by consumer business segments operating outside the United States.

For example, if the corporation's 10-K form contains information about both revenues from European operations and credit risks, we use the former to define and calculate its exposure to the European consumer market. For five of the corporations in the sample, aggregate non-U.S. revenue is the only available information about geographic distribution of the corporations' overall revenue. For these corporations, we need to additionally determine whether Europe is the corporation's primary foreign market using qualitative information in the texts in the 10-K forms (see Appendix Table C3). The European market constitute a negligible share of revenues for two of these corporations, but is the primary foreign market for the other three corporations.

Table 1 lists the names of the corporations in the sample and their exposure to the European consumer market, calculated based on public information in the annual reports (10-K forms) in 2014 – 2015 prior to the announcement of GDPR. The geographic distribution of business activities is very stable and changes little over time for these largest multinational corporations in our sample. Also, none of these corporations experienced substantial reorganization or merger events around the time that GDPR was introduced or implemented. Among sampled corporations, 10 out of 25 have below 5% exposure to GDPR, 7 out of 25 have negligible exposure to GDPR of 2% or less, and 5 out of 25 are significantly exposed with over 15% of their total revenues generated by consumer products and services provided in Europe. In the rest of the

[4] This excludes investment banking which primarily involves corporate and institutional customers, e.g., raising capital for companies, governments, and other entities, through IPO underwriting or arranging for mergers and acquisitions. It deals primarily with institutional clients; hence very limited extent of personal data is involved.

study, we apply the threshold of 5% to define the binary treatment indicator, to determine whether a corporation is *substantially exposed to GDPR*.

[Insert Table 1 here]

Most of the sampled corporations explicitly discuss GDPR in the Relevant Regulations section of their annual reports in at least one fiscal year prior to enforcement (i.e., 2017 or earlier). All the corporations that did not mention GDPR have 5% or lower exposure to the European consumer market according to our measurement approach above. To provide further validation for the measurement approach, see Appendix Section A4 for details on the measurement and validation of the GDPR exposure variable. For example, Appendix Figure B2 shows a strong positive correlation between the exposure to European consumer market and the number of times keywords related to Europe or major European countries are mentioned in the corporations' annual report.

3.4. Descriptive Statistics on Balanced Establishment-Year and Corporation-Year Panels

We conduct empirical analyses on establishments belonging to the largest U.S. financial services corporations, each with hundreds or thousands of branches located across all U.S. states. The final sample consists of a balanced panel of 86,555 observations for 17,311 establishments over five years from 2016 to 2020. To be included in the sample, the establishment must be in the finance and insurance industry (i.e., NAICS code 52) and conduct consumer-facing business activities, which can be further classified into banking and payment (522), securities and asset management (523), and insurance (524). Financial services establishments with NAICS codes indicating monetary authorities (521) and investment banking (523110) are excluded from the sample, because GDPR enforcement has little impact on them as they do not deal with individual consumers or collect personal data. Table 2 panel (a) summarizes the final sample of observations in a balanced panel from 2016 to 2020.

[Insert Table 2 here]

The average observation had a log revenue of 15.1 (or about $3.5 million US dollars). The medium and top 5% establishment spent about $0.2 million and $0.9 million on the total of IT services, hardware, and software budgets. About 35% of the establishments belong to corporations with substantial (≥5%) exposure to the European consumer market, and hence to compliance risks associated with GDPR enforcement. About 68% of the establishments belong to corporations using application programmatic interfaces (APIs) to enable internal data interoperability across subunits. The average standardization of identity and access management (IAM) software across establishments is 4% in 2017. The median establishment has 23 PCs in 2017, which is a proxy for establishment size with better measurement quality than employment size in CITDB. About 91% of the establishments in 2017 already use cloud infrastructure provided by a major public cloud vendor – Amazon, Microsoft, Google, and IBM. Almost all large enterprises rely on cloud computing technologies to deliver on-demand resources from a shared pool of distributed hardware and software across disparate environments and locations (Benlian, Kettinger, Sunyaev, & Winkler, 2018). In the years prior to GDPR enforcement, establishment revenues grew by about 3.2% on average from 2016 to 2017 in the sample.

It is worth noting that an organization may choose to increase interoperability and standardization prior to GDPR enforcement, but it requires investing significant resources and efforts into corresponding areas, and a long process that take many years to result in productive changes. The empirical correlation between revenue trend before GDPR enforcement (2016 – 2017) and GDPR exposure is very low (<0.01). If corporations made adjustments to their technology stacks or developed compliance solutions after the announcement of GDPR in 2016 but before enforcement in 2018, these changes may reflect primarily on the cost side but not on the revenue side.

Panel (b) shows the descriptive statistics for the balanced panel data underlying the estimation of the effects of GDPR enforcement on IT hiring. The sample is aggregated to the level of firm-state by year-quarter from 2016Q1 to 2019Q3, by taking the average across establishments in the CITDB and matching with the quarterly number of job postings for each corporation within a state. Job postings data are available only up to 2019Q3 and are not available at levels finer than firm-state. In panel (b), fewer than 50% of the observations contain non-zero values of the number of job postings in computer occupations. The top quartile observation posts about four jobs in computer occupations (except support specialists), among which two jobs contain regulation-related tasks, and one job is for a research scientist or analyst position. The other variables have summary statistics broadly consistent with those in panel (a).

4. Empirical Strategy

4.1. Triple Differences Regression Framework

To empirically test the hypotheses around how internal data APIs and technology component standardization affect establishment performance in response to the enforcement of the GDPR, we use the triple differences regression framework common in empirical research related to the GDPR (e.g., Burford, Shipilov, & Furr, 2022). This approach allows us to explore the heterogeneity in the treatment effect of GDPR enforcement across subsamples with different organizational characteristics. The event time is defined by GDPR enforcement in May 2018, similar to prior work assessing firm performance effects of the GDPR across various contexts (Koski & Valmari, 2020; Burford, Shipilov, & Furr, 2022; Chen, Frey, & Presidente, 2022; Johnson, Shriver, & Goldberg, 2023). Prior research has shown that more data-intensive firms are more affected by the GDPR (Jia, Jin, & Wagman, 2018).

This paper examines a relatively homogeneous set of establishments within the same sector in a single country. All the establishments are located in the United States and hence operate within similar legal frameworks and institutional contexts. All the organizations in our sample are in the financial services sector and conduct business activities that involve handling consumers' personal data. These establishments are exposed to the GDPR due to the global nature of their business segments (e.g., money transfers and payment, insurance, and banking services) that involve organizational-wide technological infrastructure for transferring data across national borders. They are likely to commit to enforcement (Koski & Valmari, 2020;

Johnson, 2022) due to potentially hefty fines charged in proportion to their global revenues, and the fact that existing processes and resources for managing regulatory compliance are already available.

To define the treatment variable, we measure the extent to which a corporation is exposed to the European consumer market, and hence subject to GDPR compliance risks. Multinational corporations with a substantial fraction of operations handling personal data of EU/EEA residents are at greater compliance risk than those with only a minor fraction of their businesses serving European customers. In Appendix Section A4, we describe in detail how we use information in corporations' annual reports (10-K forms) before 2016 to derive the GDPR exposure measure and provide various robustness checks for the validity of the measure. The exact definition of the treatment variable ($TREAT$) is an establishment-level indicator for being part of a corporation that has substantial (≥5%) exposure to the GDPR. In Equation 2 below that implements the triple differences regression design, we also use an indicator for all years since the GDPR was enforced in 2018 ($POST$), and the third variable (X) that generalizes either internal data API (API_i) or standardization of identity and access management (IAM) software (STD_{ij}). The regression controls for establishment fixed effects ϕ_{ij} and year fixed effects η_t, and hence does not additionally include variables varying at establishment level and below ($TREAT$ and $X \times TREAT$) or at the year level ($POST$). The main outcome variable Y is the log establishment revenue measured at the level of corporation i, establishment j, and year t, but it can also be other outcomes such as IT investments and IT hiring.

$$Y_{ijt} = \beta_1 TREAT_i \times POST_t + \beta_2 X_{ij} \times POST_t + \beta_3 TREAT_{ij} \times POST_t \times X_{ij} + \phi_{ij} + \eta_t + \epsilon_{ijt} \qquad (2)$$

In the triple differences regression specification of Equation 2, the coefficient β_3 identifies the role of X in inducing differential changes in the outcome variable in response to the enforcement of the GDPR. The standard errors on all the coefficient estimates need to be adjusted at the cluster (i.e., corporation) level using appropriate methods. In particular, robust inference should take into account the structure of error correlations within clusters (Cameron & Miller, 2015). Furthermore, the consistency of clustered standard error is a large-sample property which depends on a sufficiently large number of clusters. In our setting, the number of clusters is relatively small $(= 24)$. Thus, we implement the wild bootstrap method to calculate the critical values and confidence intervals through simulations, consistent with standard practice in the literature (Cameron, Gelback, & Miller, 2008; Roodman, Nielsen, MacKinnon, & Webb, 2019). This approach does not yield conventional standard errors, but it produces the p-value instead as the proportion of bootstrap test statistics more extreme than those observed in the actual sample. We report results including the point estimates, the 95% confidence intervals, and the p-values (instead of standard errors).

4.2. Matching Methods and Robustness Checks

Violation of the parallel trends assumption (PTA) is a major threat to uncovering unbiased estimates of the treatment effect in difference-in-differences regression designs (Bertrand, Duflo, & Mullainathan, 2004). Because GDPR was announced in 2016 but enforced in 2018, firms and especially large corporations may have made efforts to comply with the GDPR prior to enforcement after the regulation was announced. Evidence from previous research suggests that most companies did not begin to act on satisfying GDPR requirements until the year of enforcement in 2018 (Jia, Jin, & Wagman, 2018; Koski & Valmari, 2020; Johnson, 2022). However, large corporations might be exceptional and make early adjustments to comply because they have historically installed processes and resources for managing regulatory compliance.

We implement a few methods and robustness checks to assess the extent of parallel trends assumption (PTA) violation and to implement alternative estimation methods to address potential pre-trends. First, we use matching methods – propensity score matching (Rosenbaum & Rubin, 1983) and coarsened exact matching (Iacus, King, Porro, 2012) – to ensure that treatment and control units have similar pre-treatment characteristics. Matching is conducted on two baseline covariates that proxy for establishment size and technology intensity. These variables are log number of PCs and presence of cloud, measured in 2017, prior to GDPR enforcement. We do not match on pre-treatment outcomes to force them to be on the same trend because doing this may inject bias into the estimator (Ham & Miratrix, 2022). To assess the pre-treatment balance of covariates, we include the matched covariates and the change in log revenue from 2016 to 2017.

This allows us to check how the implemented matching affects the pre-trend in the outcome between treated and control groups. Appendix Figure B3 shows the covariate balances before and after matching.

Second, we conduct subsample analyses by estimating difference-in-differences (DID) regressions on non-overlapping subsamples split by binary indicators based on X (i.e., $\mathrm{I}(API_i = 1)$ or $\mathrm{I}\left(STD_{ij} \geq P50\right)$). We also estimate a variant of the DID regression model that replaces the single treatment effect estimate with time-varying coefficients to assess the parallel trends assumption. Third, we apply the synthetic difference-in-differences method (Arkhangelsky et al., 2021) to estimate the subsample treatment effect of GDPR enforcement on corporation-level data for internal APIs (see Appendix Section A6). The synthetic difference-in-differences method combines the strengths of both the DID and synthetic control methods, and constructs synthetic counterfactuals for the treated group by weighting control group units to minimize the distance to the treated units in pre-treatment covariates. Finally, we note that the direction of PTA violation means that firms' compliance efforts before 2018 would bias the estimated GDPR enforcement effects toward zero (Johnson, 2022), which makes our estimate the lower bound for the treatment effect in the absence of PTA violation.

5. Results

5.1. Internal Data Interoperability and Differential Effects of GDPR Enforcement

In Table 3 panel (a), we report coefficients estimated from the regression specification in Equation 2 of Section 4.1, where the regressor X is the corporation-level indicator of API which proxies for internal data interoperability. The regression is estimated on the full sample consisting of all the establishments regardless of API. The coefficient on the triple interaction ($TREAT \times POST \times API$) is 0.086 (*p=0.057*) and statistically significant at the 10% level. Using matching methods of both PSM and CEM, the triple interactions coefficient lowers to about 0.063 – 0.069 with p-values around 0.07 (columns 2 and 3). These results can be interpreted as internal data APIs mitigating the negative effects of GDPR enforcement on establishment performance by more than 6%.

[Insert Table 3 here]

These results are supported by Appendix Table C5 panel (a), which are based on alternative DID estimation on subsamples split by the binary indicator API. The estimate in column 1 shows that the enforcement of GDPR lowered the revenues of the establishments by 11.5% ($p=0.001$) among corporations without internal data APIs. In contrast, column 2 shows that the enforcement of GDPR lowered the revenues of the establishments only by about 3% ($p=0.032$) for corporations with internal data APIs. The results are based on regression specifications that control for both establishment fixed effects which absorbs all time-invariant establishment characteristics and year fixed effects.

To assess the parallel trends assumption (PTA), we plot event study regression results from subsamples split by API in Figure 1. The plots also show the 95% confidence intervals, computed using the wild bootstrap procedure which is the appropriate method for inference of treatment effects on data with few clusters (i.e., corporations) each containing a relatively large number of units (i.e., establishments). The year (t=-*1*) prior to the event (t=*0*) serves as the omitted baseline. The coefficient estimates for t=-*2* are small and the 95% confidence intervals cover zero in both panels (a) and (b). This is evidence that the parallel trends assumption (PTA) holds in both subsamples.

[Insert Figure 1 here]

In the top panel (where $API = 0$), the coefficient estimates for t=*0, 1,* and *2* indicate increasing treatment effect magnitudes over time. In the bottom panel (where $API = 1$), GDPR enforcement was followed by a small dip in revenue around 2019, which did not continue to worsen but recovered by 2020. These results suggest that adapting to regulatory compliance with GDPR is not a temporary action but can have lasting effects beyond a few years.

5.2. Technology Component Standardization and Differential Effects of GDPR Enforcement

In Table 3 panel (b), we report coefficients estimated from the regression specification in Equation 2 of Section 4.1, where the regressor X is the establishment-level continuous variable of STD measuring vendor standardization of identity and access management (IAM) software. The regression is estimated on the full sample consisting of all the establishments regardless of STD. The coefficient on the triple interaction ($TREAT \times POST \times STD$) is 4.134 ($p$=*0.046*) and statistically significant at the 5% level. Columns 2 and 3 reveal results after applying PSM and CEM to match treated and control units, which indicate very similar coefficient estimates on the triple interactions relative to column 1 with p-values less than 0.04. To interpret the results, a 1 standard deviation increase in IAM software standardization shrinks the negative effects of GDPR enforcement on establishment revenue by about 3.3% (=0.008*4.134).

These results are supported by Appendix Table C5 panel (b), which are based on alternative DID estimation on subsamples split by above and below median value (*=0.0391)* of STD. The estimate in column 1 shows that the enforcement of GDPR lowered the revenues of the establishments by 8.7% (p=*0.034*) among corporations with below median IAM software standardization. In contrast, column 2 shows that the enforcement of GDPR had a negligible impact on revenue for corporations with above median IAM software standardization, with a coefficient estimate of *0.004* and substantial variance (p=*0.118*). The subsamples consist of establishments with at least one product vendor present in 2017 in the relevant category, so that STD is well-defined. All the regression specifications control for both establishment fixed effects which absorbs all time-invariant establishment characteristics and year fixed effects which account for time trends.

Figures 2 shows the event study regression results by plotting coefficient estimates separately for each year relative to the year of GDPR enforcement, which allows us to assess the parallel trends assumption (PTA). The samples underlying the top and bottom panels are establishments with below median and above median STD, respectively. 95% confidence intervals are computed using the wild bootstrap procedure which is the appropriate method for inference of treatment effects on data with few clusters (i.e., corporations) each containing a relatively large number of units (i.e., establishments). The year (*t=-1*) prior to the event (*t=0*) serves as the omitted baseline. The coefficient estimates for *t=-2* are very close to zero, and the 95% confidence intervals cover zero in both subfigures, which lends evidence for the parallel trends assumption (PTA) in both subsamples.

[Insert Figure 2 here]

Panels (a) and (b) focus on the subsamples with below-median and above-median standardization in identity and access management software relative to all establishments, respectively. In the top panel (where $STD < P50$), the coefficient estimates for *t=0, 1,* and *2* indicate increasing treatment effect magnitudes over time. In the bottom panel (where $STD \geq P50$), GDPR enforcement was followed by a small dip in revenue around 2019 which recovered by 2020. These estimates are very noisy and have wide 95% confidence intervals. These results suggest that the negative impact of GDPR enforcement on establishment revenues was not transitory for data-intensive corporations studied in this setting. However, the effects of GDPR enforcement are heterogeneous across establishments' level of technology component standardization. Establishments that use highly standardized IAM software do not experience significant performance decline, while those using uncommon or novel versions of the software suffer substantial revenue decline.

5.3. IT Investments, Labor Adjustments, and Compliance Costs

In this section, we explore channels of adaptation and firms' actions in response to the enforcement of the GDPR. We focus on two sets of outcome variables – IT investments and IT hiring. First, changes in IT investments reflect the extent to which firms engage in value creation enabled by data and digital technologies. A decline in IT investments suggests a reduction in value-enhancing uses of IT services and products, which can lower firms' revenue performance. We estimate the impact of the GDPR enforcement at the establishment-year level on the IT investment outcomes (from 2016 to 2020). Second, changes in IT hiring reflects firms' efforts to comply with the regulation, by hiring workers with technical expertise to develop compliance solutions and perform regulation-related tasks. We estimate the impact of the GDPR enforcement at the firm-state-quarter level on the IT hiring outcomes from 2016Q1 to 2019Q3. For regressions with IT hiring outcome variables, the time of treatment is 2018Q2, which contains May 2018 when the GDPR was implemented. The number of time periods relative to treatment in an event study is thus defined by the number of quarters that have elapsed since (or before) 2018Q2.

Table 4 shows how corporations change their levels of IT investments and IT hiring in response to GDPR enforcement, conditional on the availability of internal data APIs. We use the difference-in-differences estimation framework similar to Equation 2 in Section 4.1 to estimate the treatment effect of having substantial (≥5%) exposure to the European consumer market on IT investment and IT hiring. For corporations without internal data APIs, the regression results in panel (a) show that GDPR enforcement led to an 8 – 9% decline in establishment annual budgets for IT investments (measured by total IT services, hardware, and software). These effects are highly significant at the 1% level according to the confidence intervals and p-values calculated based on the wild bootstrap method, which is the appropriate method for statistical inference on data structures with a small number of clusters. On the other hand, regression results in panel (b) show that the corporations without data APIs affected by the GDPR post 80% more IT job, significant at the 1% level (column 1). Similar results hold for both IT research scientists and analysts (column 3) and IT occupations requiring regulation-related tasks (column 5). In contrast, corporations with data APIs do not experience significant changes in either IT investments or job postings in response to GDPR enforcement.

[Insert Table 4 here]

These results indicate that the lack of internal data interoperability forces corporations to respond to requirements for GDPR compliance by downscaling investments in value-enhancing digital technologies and increasing the size of its IT workforce. GDPR enforcement requires corporations to have the ability to track internal data. Firms can orchestrate data flows either through automated or manual interfaces. While APIs make it relatively easy to facilitate data interoperability at a low cost, they also require coordination and centralized decisions to develop and maintain for which some organizations are not ready. To meet regulatory requirements, firms may remove or scale down IT services and products that risk compliance violation. On the other hand, IT workers specializing in data-intensive tasks (i.e., analysts and research scientists) may be an imperfect substitute for automated interfaces and firms need more of them to implement compliance solutions when the necessary data architecture for automating data flows is not ready. Similarly, firms increased their demand for workers carrying out tasks to ensure regulatory compliance.

Figure 3 plots the event study regression results by showing separate estimates for each time period relative to the event. The left panel (a) focuses on total IT investments (including IT services, hardware, and software) and the right panel (b) focuses on IT job postings (except support specialists) as the outcome variables. Both panels report coefficient estimates along with 95% confidence intervals for all time periods (except *t=-1*). The graph shows that the differences in outcomes across treatment and control groups are statistically insignificant and close to zero prior to 2018 (or 2018Q2) between establishments (or firm-state pairs). This is evidence for that the parallel trends assumption (PTA) is satisfied. The event plots show that the downscaling of IT investments and increase in IT hiring are not temporary but intensify over time.

[Insert Figure 3 here]

Table 5 shows how corporations change their levels of IT investments and IT hiring in response to GDPR enforcement, conditional on technology component standardization of identity and access management (IAM) software. We estimate the effects of GDPR enforcement on regional (state-level) IT hiring by the focal corporation in a given year-quarter relative to 2018Q2, using the difference-in-differences estimation framework similar to Equation 2 in Section 4.1. The regression results in panel (a) suggest that for corporations with below-median IAM vendor standardization, GDPR enforcement led to about 4% decline of establishment annual budgets for IT services, hardware, and software, significant at the 5% level. The regression results in panel (b) show that these corporations did not increase or decrease their hiring of IT workers significantly. These results indicate that that lack of standardization made complying with the regulation more difficult. However, firms can choose to remove non-compliant IT products to mitigate risks of violating the GDPR, which contributes to lower IT spending since 2018.

[Insert Table 5 here]

To summarize, the empirical analyses in this section shed light on the mechanisms of adjustment by corporations that are substantially exposed to the GDPR. The results imply that lowering IT investments into value-creating IT services and products may be an important channel explaining the revenue decline after GDPR enforcement among corporations without data APIs and with low vendor standardization technologies affected by the regulation. These results provide further evidence that architectural factors that facilitate interoperability in data and applications may remain the bottleneck to value creation from data and are crucial to improving resilience to regulatory shocks.

6. Discussion and Conclusion

In this paper, we study the impact of two features of the technological architecture – internal data interoperability and software standardization – on organizational adaptation in response to system-wide regulatory shock that requires stricter internal data governance. Interoperability and standardization contribute to internal data governance and thus help organizations adapt effectively to the Global Data Privacy Regulation (GDPR) enforcement. Our results show that internal data APIs and identity and access management (IAM) software standardization significantly mitigate establishments' performance decline following GDPR enforcement. The performance decline is likely caused by lowering value-enhancing IT investments to comply with the regulation, as we find evidence that GDPR enforcement lowered the IT budget among corporations with low interoperability and standardization. On the side of compliance costs, empirical evidence suggests that corporations without internal data APIs substantially increased hiring efforts in computer occupations, especially in data-intensive and regulation-related roles. Thus, corporations with low internal data interoperability incur high costs to comply with the regulation.

Data governance problems in large organizations are challenging because they require joint coordination across many actors (e.g., Benfeldt, Persson, & Madsen, 2020). For legacy corporations, internal data governance is essential for regulatory compliance and digital innovation objectives, but it requires intraorganizational coordination across multiple functional units. Integrative capability (Helfat & Campo-Rembado, 2016; Helfat & Raubitschek, 2018) can create conditions that remove technological bottlenecks for data governance, but it remains somewhat neglected by the digital innovation literature that predominantly associates value creation from data and digital technologies with decentralized architecture and distributed systems. Our results show that interoperability and standardization, which require collective decisions beyond local units, help organizations adapt to GDPR enforcement effectively. They increase the agility with which large corporations can maintain a holistic internal view of all their data sources and deploy compliance solutions at a global scale to meet regulatory requirements. The qualitative data we collect from online news articles and public interviews with senior technology executives in financial services corporations align broadly with these interpretations of the quantitative results.

Data interoperability has received much attention from regulators and policymakers as a potential mandate to reign in the market power of large corporations and facilitate industry competition (e.g., Martens, Parker, Petropoulos, & Van Alstyne, 2021; Bourreau, Krämer, & Buiten, 2022). The UK Open Banking regulation specifically mandated regulatory technical standards for data interoperability in the financial services sector (Dinckol, Ozcan, & Zachariadis, 2023). However, the implementation was complex and did not yield the intended results due to substantial variations in incentives, architecture, and technical capabilities across industry players. Incumbent legacy firms in the established sector constitute a vital context in a complex reality that can make implementing mandated regulatory standards difficult. On the other hand, standardizing technology components can lower the barriers to achieving interoperability, which also increases concentration in the supplier market and discourages novel solutions that depart substantially from industry standards (e.g., Miric, Ozalp, & Yilmaz, 2023), which are the opposite of regulators' objectives for fostering competition and innovation.

This paper makes several contributions. First, it provides quantitative evidence on the technological features that ensure internal data governance. While the theoretical arguments for the benefits of joint coordination and integrative capability have been made (Helfat & Raubitschek, 2018; Melville & Kohli, 2021; Widjaja & Gregory, 2020), there has been little empirical work measuring integrative capability or illustrating how they affect performance outcomes and adaptation costs of large corporations in response to changes in the external environment. Our results add to the understanding of technological conditions for facilitating internal data governance and responding to changing regulatory mandates.

We also contribute to the literature on the effects of privacy regulation, particularly the General Data Protection Regulation (GDPR). Existing research reveals the impact of GDPR on firm behavior across different contexts (Johnson, 2022; Johnson, Shriver, & Goldberg, 2023; Wang, Jiang, & Yang, 2023; Peukert, Bechtold, Batikas, & Kretschmer, 2022; Godinho de Matos & Adjerid, 2022; Burford, Shipilov, & Furr, 2022; Chen, Frey, & Presidente, 2022; Zhuo, Huffaker, & Greenstein, 2021; Koski & Valmari, 2020; Gal & Aviv, 2020; Martin, Matt, Niebel, & Blind, 2019; Jia, Jin, & Wagman, 2018). However, most existing literature focused on advertising and consumer-level outcomes and identified short-run effects that indicate substantial heterogeneity in compliance efforts across markets and firms. Little is known about how GDPR enforcement affected organizational strategies and performance beyond the short run among large data-intensive corporations. Our findings suggest that interoperability and standardization led to substantial differences in compliance adjustments and revenues following GDPR enforcement, and the effects persisted over time.

Our study offers a few managerial implications and practical recommendations. Digital technologies have become a source of dynamic disruption, introducing new business and operating models that significantly depart from those of existing industry incumbents (e.g., Eklund & Kapoor, 2019). While incumbent corporations may embark on digital transformation to adopt new business models, these transformation programs take a long time to implement and fail to yield the intended benefits. We show that integrative capability can be a valuable goal of transformation, which requires joint coordination of intraorganizational actors and different functional units. A global approach may make it easier for organizations to achieve this goal and facilitate digital value creation by properly orchestrating inter-system linkages and resource sharing. Such an approach requires mechanisms beyond localized innovation and fragmented technology solutions. For example, local owners of APIs need to acknowledge that users from other business units will access the APIs they developed, and standardization requires multiple sub-units to agree upon the shared design and implementation of standard technology components.

Our results are also relevant to policymakers designing privacy regulations to limit societal risks of personal data exploitation. The GDPR was a pioneering regulatory framework for personal data privacy protection, upon which more recent regulatory efforts have emerged. Understanding the intended and unintended consequences of GDPR enforcement is crucial, as is identifying the technological features that can enable or hinder organizational adaptation. The incentives to standardize technology vendor choices may stifle competition in the supplier market and reduce the variety of local experimentation and data-driven innovation. We point out these potential tradeoffs that involve constraints on the technology choices of large corporations to meet privacy regulation mandates. Finally, value creation mechanisms based on data and digital technologies may vary across sectors. Future studies can examine how different technological features facilitate data governance across sectors with different regulatory and market environments.

References

Agrawal, A., Gans, J. S., & Goldfarb, A. (2023). Artificial Intelligence Adoption and System-Side Change. *Journal of Economics & Management Strategy*.

Aghion, P., Bloom, N., Lucking, B., Sadun, R., & Van Reenen, J. (2021). Turbulence, Firm Decentralization, and Growth in Bad Times. *American Economic Journal: Applied Economics*, *13*(1), 133–169.

Albert, D., & Siggelkow, N. (2022). Architectural Search and Innovation. *Organization Science*, *33*(1), 275–292.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic Difference-in-Differences. *American Economic Review*, *111*(12), 4088–4118.

Baldwin, C. Y., & Clark, K. B. (2000). *Design Rules: The Power of Modularity* (Vol. 1). MIT Press.

Barrett, M., Davidson, E., Prabhu, J., & Vargo, S. L. (2015). Service Innovation in the Digital Age: Key Contributions and Future Directions. *MIS Quarterly*, *39*(1), 135–154.

Benfeldt, O., Persson, J. S., & Madsen, S. (2020). Data Governance as a Collective Action Problem. *Information Systems Frontiers*, *22*, 299–313.

Benlian, A., Kettinger, W. J., Sunyaev, A., & Winkler, T. J. (2018). The Transformative Value of Cloud Computing: A Decoupling, Platformization, and Recombination Theoretical Framework. *Journal of Management Information Systems*, *35*(3), 719–739.

Benzell, S., Hersh, J. S., & Van Alstyne, M. W. (2022). How APIs Create Growth by Inverting the Firm. *Management Science*.

Berman, R., & Israeli, A. (2022). The Value of Descriptive Analytics: Evidence from Online Retailers. *Marketing Science*, *41*(6), 1074–1096.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275.

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

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N. V. (2013). Digital Business Strategy: Toward a Next Generation of Insights. *MIS Quarterly*, 471–482.

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

Boh, W. F., & Yellin, D. (2006). Using Enterprise Architecture Standards in Managing Information Technology. *Journal of Management Information Systems*, *23*(3), 163–207.

Boroomand, F., Leiponen, A., & Vasudeva, G. (2022). Does the Market Value Attention to Data Privacy? Evidence from US-Listed Firms Under the GDPR. *Wharton Mack Institute Working Paper*.

Bourreau, M., Krämer, J., & Buiten, M. (2022). *Interoperability in Digital Markets.*

Burford, N., Shipilov, A. V., & Furr, N. R. (2022). How Ecosystem Structure Affects Firm Performance in Response to a Negative Shock to Interdependencies. *Strategic Management Journal*, *43*(1), 30–57.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, *90*(3), 414–427.

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372.

Campagnolo, D., & Camuffo, A. (2010). The Concept of Modularity in Management Studies: A Literature Review. *International Journal of Management Reviews*, *12*(3), 259–283.

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

Chwelos, P., Ramirez, R., Kraemer, K. L., & Melville, N. P. (2010). Research Note—Does Technological Progress Alter the Nature of Information Technology as a Production Input? New Evidence and New Results. *Information Systems Research*, *21*(2), 392–408.

Dallemule, L., & Davenport, T. H. (2017). What's Your Data Strategy. *Harvard Business Review*, *95*(3), 112–121.

Dinckol, D., Ozcan, P., & Zachariadis, M. (2023). Regulatory Standards and Consequences for Industry Architecture: The Case of UK Open Banking. *Research Policy*, *52*(6), 104760.

Drechsler, K., Gregory, R. W., Wagner, H.-T., & Tumbas, S. (2020). At the Crossroads between Digital Innovation and Digital Transformation. *Communications of the Association for Information Systems.*, *47*(1), 23.

Eklund, J., & Kapoor, R. (2019). Pursuing the New While Sustaining the Current: Incumbent Strategies and Firm Value During the Nascent Period of Industry Change. *Organization Science*, *30*(2), 383–404.

Elm, M., Gaughan, M., & Brown, T. (2021). *The Banking Heads of Digital Report*. Insider Intelligence. https://www.insiderintelligence.com/insights/bank-of-america-head-of-digital-david-tyrie-interview/

Englmaier, F., Galdon-Sanchez, J. E., Gil, R., & Kaiser, M. (2019). *Management Practices and Firm Performance During the Great Recession*.

Ethiraj, S. K., & Levinthal, D. (2004). Modularity and Innovation in Complex Systems. *Management Science*, *50*(2), 159–173.

Faulkner, P., & Runde, J. (2019). Theorizing the Digital Object. *MIS Quarterly*, *43*(4).

Forman, C., Goldfarb, A., & Greenstein, S. (2005). How Did Location Affect Adoption of the Commercial Internet? Global Village vs. Urban Leadership. *Journal of Urban Economics*, *58*(3), 389–420.

Gal, M. S., & Aviv, O. (2020). The Competitive Effects of the GDPR. *Journal of Competition Law & Economics*, *16*(3), 349–391.

Giustiziero, G., Kretschmer, T., Somaya, D., & Wu, B. (2022). Hyperspecialization and Hyperscaling: A Resource-Based Theory of the Digital Firm. *Strategic Management Journal*.

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

Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The Role of Artificial Intelligence and Data Network Effects for Creating User Value. *Academy of Management Review*, *46*(3), 534–551.

Gregory, R. W., Keil, M., & Muntermann, J. (2012). *Ambidextrous IS Strategy: The Dynamic Balancing Act of Developing a 'Transform & Merge' Strategy in the Banking Industry.*

Ham, D. W., & Miratrix, L. (2022). *Benefits and Costs of Matching Prior to a Difference in Differences Analysis When Parallel Trends Does Not Hold*.

Helfat, C. E., & Campo-Rembado, M. A. (2016). Integrative Capabilities, Vertical Integration, and Innovation Over Successive Technology Lifecycles. *Organization Science*, *27*(2), 249–264.

Helfat, C. E., & Raubitschek, R. S. (2018). Dynamic and Integrative Capabilities for Profiting from Innovation in Digital Platform-Based Ecosystems. *Research Policy*, *47*(8), 1391–1399.

Henderson, R. M., & Clark, K. B. (1990). Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. *Administrative Science Quarterly*, *35*(9–30).

High, P. (2023). *Former Amazon Exec Marco Argenti Drives A Remarkable Digital Transformation At Goldman Sachs*. Forbes. https://www.forbes.com/sites/peterhigh/2023/01/25/former-amazon-exec-marco-argenti-drives-a-remarkable-digital-transformation-at-goldman-sachs/

Huang, P., Ceccagnoli, M., Forman, C., & Wu, D. J. (2022). IT Knowledge Spillovers, Absorptive Capacity, and Productivity: Evidence from Enterprise Software. *Information Systems Research*, *33*(3), 908–934.

Iacus, S. M., King, G., & Porro, G. (2012). Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political Analysis*, *20*(1), 1–24.

Jia, J., Jin, G. Z., & Wagman, L. (2018). The Short-Run Effects of GDPR on Technology Venture Investment. *National Bureau of Economic Research*, *w25248*.

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

Johnson, G. A., Shriver, S. K., & Goldberg, S. G. (2023). Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR. *Management Science*.

Kaganer, E., Gregory, R. W., & Sarker, S. (2023). A Process for Managing Digital Transformation: An Organizational Inertia Perspective. *Journal of the Association for Information Systems*, *24*(4), 1005–1030.

Kolbjornsen, C., & Rockwood, H. (2019). *Wells Fargo Launches New Brand Campaign, 'This is Wells Fargo,' Focused on Customer Experience*. Business Wire.

https://www.businesswire.com/news/home/20190124005671/en/Wells-Fargo-Launches-New-Brand-Campaign-'This-is-Wells-Fargo'-Focused-on-Customer-Experience

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

Kretschmer, T., & Khashabi, P. (2020). Digital Transformation and Organization Design: An Integrated Approach. *California Management Review*, *62*(4), 86–104.

Kretschmer, T., Miravete, E. J., & Pernías, J. C. (2012). Competitive Pressure and the Adoption of Complementary Innovations. *American Economic Review*, *102*(4), 1540–1570.

Langlois, R. N., & Robertson, P. L. (1992). Networks and Innovation in a Modular System: Lessons from the Microcomputer and Stereo Component Industries. *Research Policy*, *21*(4), 297–313.

Levy, M. (2015). *Freed from Harte-Hanks, the AccessCI Database May Become Relevant Again.* https://gzconsulting.org/2015/10/03/freed-from-harte-hanks-the-accessci-database-may-become-relevant-again/

Levy, M. (2019). *Aberdeen Behavioral Technographics.* https://gzconsulting.org/2019/10/07/aberdeen-behavioral-technographics/

Lippert, M., & Kane, G. (2017). *MetLife Centers Its Strategy on Digital Transformation*. MIT Sloan Management Review. https://sloanreview.mit.edu/article/metlife-centers-its-strategy-on-digital-transformation/

Loebbecke, C., & Picot, A. (2015). Reflections on Societal and Business Model Transformation Arising from Digitization and Big Data Analytics: A Research Agenda. *The Journal of Strategic Information Systems*, *24*(3), 149–157.

Martens, B., Parker, G., Petropoulos, G., & Van Alstyne, M. W. (2021). Towards Efficient Information Sharing in Network Markets. *Working Paper*.

Martin, N., Matt, C., Niebel, C., & Blind, K. (2019). How Data Protection Regulation Affects Startup Innovation. *Information Systems Frontiers*, *21*, 1307–1324.

McElheran, K. (2014). Delegation in Multi-Establishment Firms: Adaptation vs. Coordination in I.T. Purchasing Authority. *Journal of Economics and Management Strategy*, *2*, 225–257.

Melville, N. P., & Kohli, R. (2021). Models for API Value Generation. *MIS Quarterly Executive*, *20*(2).

Miric, M., Ozalp, H., & Yilmaz, E. D. (2023). Trade-Offs to Using Standardized Tools: Innovation Enablers or Creativity Constraints? *Strategic Management Journal*, *44*(4), 909–942.

Mithas, S., Tafti, A., & Mitchell, W. (2013). How a Firm's Competitive Environment and Digital Strategic Posture Influence Digital Business Strategy. *MIS Quarterly*, *37*(2), 511–536.

Morgan Stanley Infrastructure Partners: Why the Future is Digital. (2021). Infrastructure Investor. https://www.infrastructureinvestor.com/morgan-stanley-infrastructures-partners-why-the-future-is-digital/

Nagle, F. (2019). Open Source Software and Firm Productivity. *Management Science*, *65*(3), 1191–1215.

Noyes, K. (2021). *Prudential Banks on Transformation to Ensure Its Future*. The Wall Street Journal: CIO Journal. https://deloitte.wsj.com/articles/prudential-banks-on-transformation-to-ensure-its-future-01619809331?tesla=y&tesla=y

Peukert, C., Bechtold, S., Batikas, M., & Kretschmer, T. (2022). Regulatory Spillovers and Data Governance: Evidence from the GDPR. *Marketing Science*, *41*(4), 746–768.

Piccoli, G., Rodriguez, J., & Grover, V. (2023). Digital Strategic Initiatives and Digital Resources: Construct Definition and Future Research Directions. *MIS Quarterly*, *46*(4), 2289–2316.

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and Wild: Bootstrap Inference in Stata Using boottest. *The Stata Journal*, *19*(1), 4–60.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55.

Sambhara, C., Rai, A., & Xu, S. X. (2022). Configuring the Enterprise Systems Portfolio: The Role of Information Risk. *Information Systems Research*, *33*(2), 446–463.

Sanchez, R., & Mahoney, J. T. (1996). Modularity, Flexibility, and Knowledge Management in Product and Organization Design. *Strategic Management Journal*, *17*(S2), 63–76.

Simon, H. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, *106*(6), 467–482.

Teece, D. J. (2007). Explicating Dynamic Capabilities: The Nature and Microfoundations of (Sustainable) Enterprise Performance. *Strategic Management Journal*, *28*(13), 1319–1350.

The CIO's role in digital transformation. (2017). Optum. https://www.optum.com/content/dam/optum3/optum/en/resources/articles-blog-posts/WF495098-cios-role-digital-transformation-article.pdf

The Digital Transformation: Speed and convenience drive B2C payments. (2016). J.P. Morgan. https://www.jpmorgan.com/solutions/treasury-payments/insights/digital-transformation

The evolution of MoneyGram: A digital transformation success story. (2021). Business Reporter. https://www.business-reporter.co.uk/finance/the-evolution-of-moneygram-a-digital-transformation-success-story

Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Research Commentary—Digital Infrastructures: The Missing IS Research Agenda. *Information Systems Research*, *21*(4), 748–759.

Trebbi, F., Zhang, M. B., & Simkovic, M. (2023). *The Cost of Regulatory Compliance in the United States.*

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

Ulrich, K. (1995). The Role of Product Architecture in the Manufacturing Firm. *Research Policy*, *24*(3), 419–440.

Vial, G. (2019). Understanding Digital Transformation: A Review and a Research Agenda. *The Journal of Strategic Information Systems*, *28*(2), 118–144.

Wang, P., Jiang, L., & Yang, J. (2023). The Early Impact of GDPR Compliance on Display Advertising: The Case of an Ad Publisher. *Journal of Marketing Research.*

Wessel, L., Baiyere, A., Ologeanu-Taddei, R., Cha, J., & Blegind-Jensen, T. (2021). Unpacking the Difference Between Digital Transformation and IT-Enabled Organizational Transformation. *Journal of the Association for Information Systems*, *22*(1), 102–129.

Widjaja, T., & Gregory, R. W. (2012). Design Principles for Heterogeneity Decisions in Enterprise Architecture Management. *33rd International Conference on Information Systems*.

Wu, L., Hitt, L., & Lou, B. (2020). Data Analytics, Innovation, and Firm Productivity. *Management Science*, *66*(5), 2017–2039.

Xia, W., & Lee, G. (2005). Complexity of Information Systems Development Projects: Conceptualization and Measurement Development. *Journal of Management Information Systems*, *22*(1), 45–83.

Zakerinia, S., Yang, N., & Rao, V. R. (2023). Strategic Modular Innovation. *Working Paper*.

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

Figure 1: Effects of GDPR Enforcement on Revenue by Internal Data APIs. The plotted estimates are the treatment effects and 95% confidence intervals of GDPR enforcement over time, based on an event-study specification estimating the impact of GDPR enforcement on establishment revenue. The year before the event serves as the excluded baseline. Controls include establishment fixed effects and year fixed effects. Confidence intervals are estimated using wild cluster bootstrap with N=9999, which provides more conservative CI estimates than cluster robust standard errors when the data structure consists of a small number of large clusters. The subfigures compare establishments across corporations responding "No" (top) and "Yes" (bottom) to the survey question: *"Do you use services interfaces or APIs to publish departmental (e.g., not application) information between departments and/or functional areas?"*

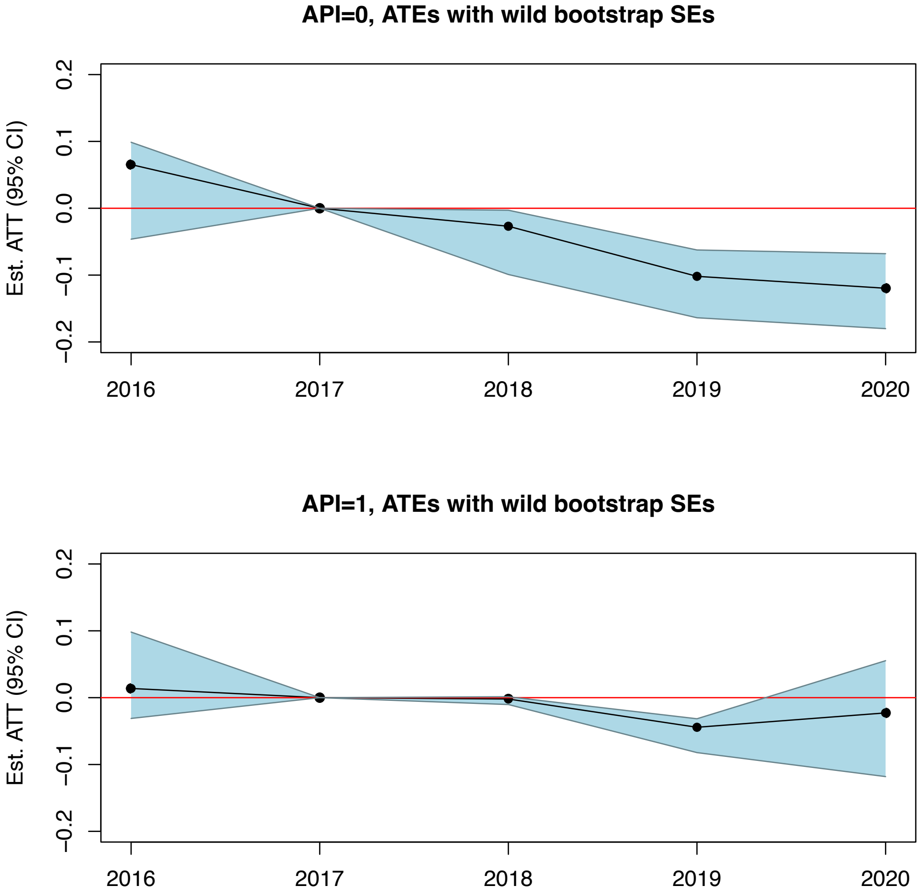
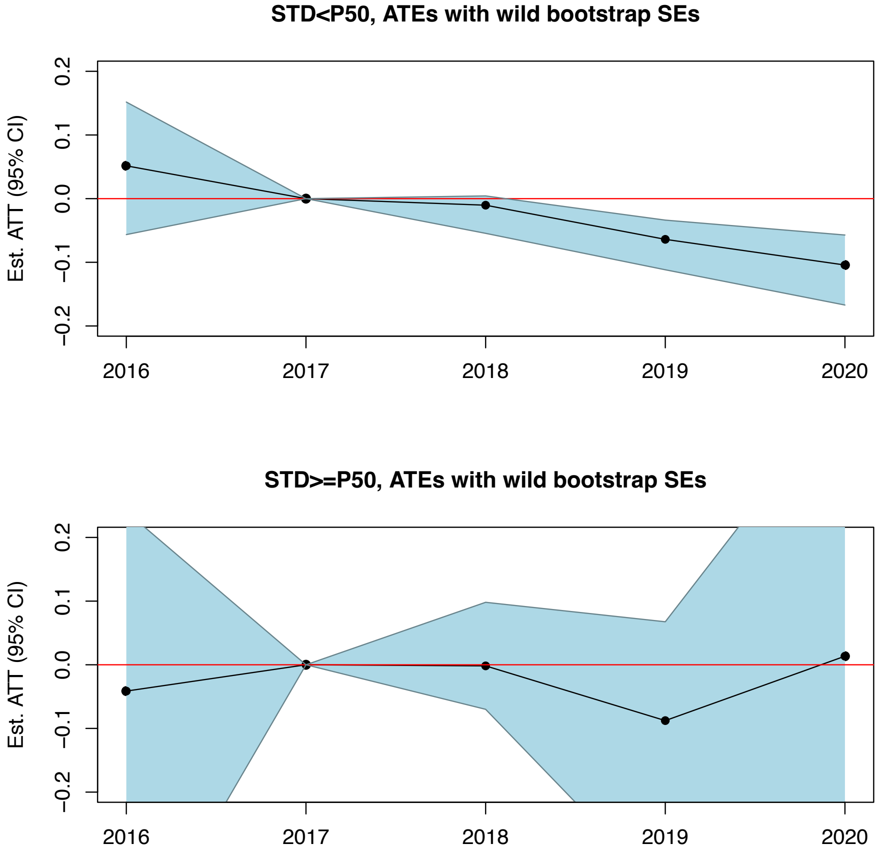


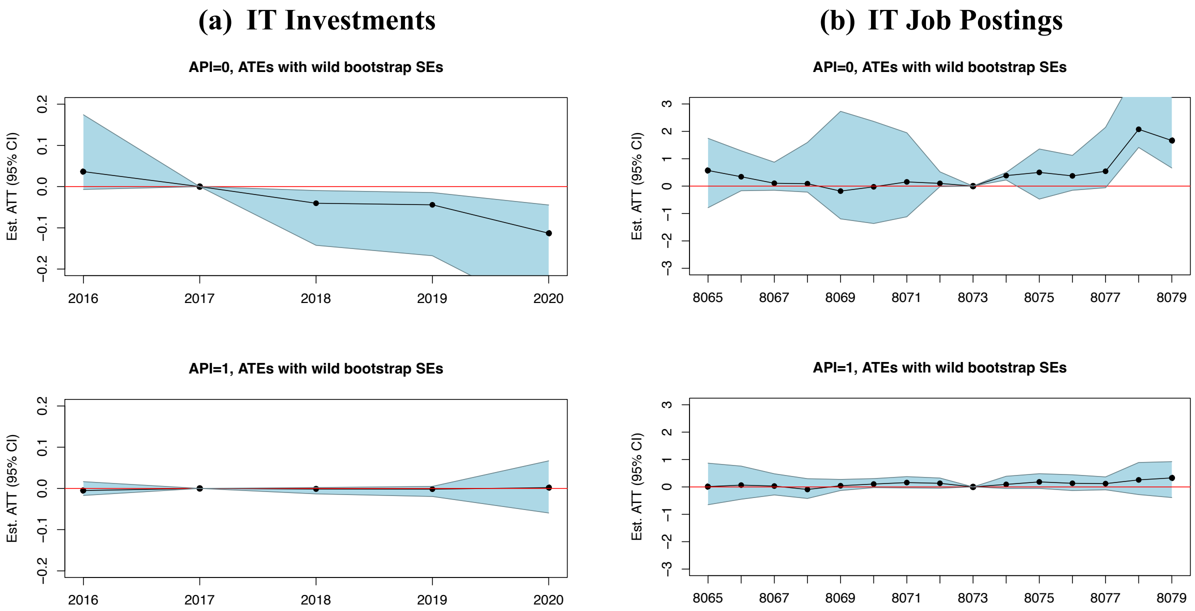
Figure 2: Effects of GDPR Enforcement on Revenue by Standardization of IAM Software. The plotted estimates are the treatment effects and 95% confidence intervals of GDPR enforcement over time, based on an event-study specification estimating the impact of GDPR enforcement on establishment revenue. The year before the event serves as the excluded baseline. Controls include establishment fixed effects and year fixed effects. Confidence intervals are estimated using wild cluster bootstrap with N=9999, which provides more conservative CI estimates than cluster robust standard errors when the data structure consists of a small number of large clusters. The subfigures compare establishments with identity and access management software standardization below (top) and above (bottom) median values.



STD<P50, ATEs with wild bootstrap SEs

STD>=P50, ATEs with wild bootstrap SEs

Figure 3: Effects of GDPR Enforcement on IT Investments and IT Job Postings by Internal Data APIs. The plotted estimates are the treatment effects and 95% confidence intervals of GDPR enforcement over time, based on an event-study specification estimating the impact of GDPR enforcement on log IT investments in panel (a) and IT job postings in panel (b). In panel (a), the outcome is log IT investments (calculated as the sum of IT services, hardware, and software budgets). In panel (b), the outcome is log job postings for IT occupations (excluding user support specialists). The year before the event serves as the excluded baseline. Controls include establishment fixed effects and year fixed effects. Confidence intervals are estimated using wild cluster bootstrap with N=9999, which provides more conservative CI estimates than cluster robust standard errors when the data structure consists of a small number of large clusters. The subfigures compare establishments across corporations responding "No" (top) and "Yes" (bottom) to the survey question: *"Do you use services interfaces or APIs to publish departmental (e.g., not application) information between departments and/or functional areas?"*



(a) IT Investments

(b) IT Job Postings

Table 1: Corporations' Exposure to the European Consumer Market. This table summarizes the estimated share of revenue (2014-15) exposed to the European market (in all business segments except investment banking). Column 3 contains the information source for determining exposure. Column 4 lists each corporation's major business segments involving the intensive collection and use of personal data.

Company Name	Exposure	Source	Business Segments
MoneyGram International, Inc.	45%	4a	Payment
Mastercard Incorporated	39%	4a	Payment
The Bank of New York Mellon Corporation	26%	1	Asset Management
The Goldman Sachs Group, Inc.	20%	1	Asset Management
Chubb Limited	16%	3	Insurance
Citigroup Inc.	13%	1	Consumer Banking, Payment
Morgan Stanley	12%	1	Asset Management
American Express Company	10%	1	Payment
American International Group, Inc.	10%	1	Insurance
Markel Corporation	10%	2	Insurance
Bank of America Corporation	7%	1	Consumer Banking, Payment, Asset Management
Assurant, Inc.	7%	2	Insurance
The Hartford Financial Services Group, Inc.	6%	2	Insurance
MetLife, Inc.	5%	1	Insurance
Eaton Vance Corp.	5%	4a	Asset Management
JPMorgan Chase & Co.	4%	1	Consumer Banking, Payment, Asset Management
Fifth Third Bancorp	3%	2	Consumer Banking
Wells Fargo & Company	3%	2	Consumer Banking, Asset Management
Capital One Financial Corporation	2%	1	Consumer Banking, Payment
The Travelers Companies, Inc.	2%	1	Insurance
Lincoln National Corporation	0%	1	Insurance
Centene Corporation	0%	1	Insurance
The Charles Schwab Corporation	0%	2	Asset Management
Prudential Financial, Inc.	0%	4b	Insurance, Asset Management
UnitedHealth Group Incorporated	0%	4b	Insurance

Notes:

(1) Codes for information source: [1] EMEA (Europe, the Middle East and Africa) revenue share [2] credit risk exposure to European countries [3] net premiums earned for insurance companies [4a] non-US revenue share when Europe is the primary foreign market [4b] zero when information on revenue by geography is only available for aggregate non-US and operations in Europe have insignificant contribution to revenue. Credit risk exposure is calculated as total credit exposure to European countries divided by total interest earning assets. See Appendix Section A4, Appendix Figure B2, and Appendix Table C3 for further details.

(2) Revenues from investment banking are excluded from calculating the revenue shares for determining exposure to the European consumer market. This is because investment banking deals with institutional (non-personal) clients that are unaffected by the GDPR which regulates the usage and processing of personal data. Only three corporations are affected by this adjustment: JPMorgan Chase, Goldman Sachs, and Bank of America.

(3) All except five corporations mention GDPR explicitly in their 10-Ks' Regulation section prior to 2018. The only corporations that did not mention GDPR are Eaton Vance Corp, Fifth Third Bancorp, Wells Fargo, Centene Corporation, Lincoln National Corporation, and Charles Schwab. All of them had 5% or less exposure to the European consumer market.

Table 2: Descriptive Statistics. Panel (a) summarizes the balanced panel at the establishment-year level. Panel (b) summarizes the balanced panel at the firm-state-year level. In panel (a), the total IT spend is calculated as the sum of IT services, hardware, and software budgets. In panel (b), the total number of IT job postings is calculated as the sum of the three subcategories of IT job postings. The 2010 SOC occupation codes for IT-Info jobs are 15-1111, 15-1121, and 15-1122 (Computer and Information Research Scientists & Analysts). The 2010 SOC occupation codes for IT-Net jobs are 15-1141, 15-1142, and 15-1143 (Database & Systems Admin & Network Architects). The 2010 SOC occupation codes for IT-Dev jobs are 15-1169, 15-1170, 15-1171, and 15-1172 (Software Developers & Programmers).

(a) Establishment Level

Variable	Mean	SD	P5	P25	P50	P75	P95	# Obs.
Ln(Revenue)	15.063	1.128	13.814	14.509	14.914	15.269	17.148	86555
Ln(Total IT Spend)	12.053	1.004	10.951	11.359	12.056	12.293	13.697	86555
Ln(IT Services Spend)	11.274	0.991	10.100	10.602	11.299	11.536	12.916	86555
Ln(Hardware Spend)	9.802	0.998	8.700	9.118	9.816	10.052	11.440	86555
Ln(Software Spend)	11.215	1.021	10.086	10.502	11.199	11.440	12.918	86555
GDPR Exposure ≥5%	0.354	0.478	0.000	0.000	0.000	1.000	1.000	86555
Internal Data APIs	0.681	0.466	0.000	0.000	1.000	1.000	1.000	86555
Standardization IAM	0.037	0.008	0.024	0.034	0.039	0.046	0.047	65740
Log #PCs	3.149	0.757	2.197	2.708	3.135	3.332	4.407	86555
Cloud Presence	0.914	0.280	0.000	1.000	1.000	1.000	1.000	86555
ΔRevenue (2016-17)	0.032	0.122	0.001	0.001	0.041	0.043	0.114	86555

(b) Firm-State Level

Variable	Mean	SD	P5	P25	P50	P75	P95	# Obs.
Ln(#IT Job Postings)	0.961	1.529	0.000	0.000	0.000	1.609	4.494	5640
Ln(#IT-Info Job Postings)	0.489	1.025	0.000	0.000	0.000	0.693	2.944	5640
Ln(#IT-Reg Job Postings)	0.754	1.313	0.000	0.000	0.000	1.099	3.795	5640
GDPR Exposure ≥5%	0.527	0.499	0.000	0.000	1.000	1.000	1.000	5640
Internal Data APIs	0.686	0.464	0.000	0.000	1.000	1.000	1.000	5640
Standardization IAM	0.033	0.012	0.018	0.024	0.039	0.046	0.047	3285

Table 3: Triple Differences Regression Estimates. The tables show regression results estimating the triple differences regression specification of Equation 2 in Section 4.1 on establishment-year panel data. In panel (a), the third variable is internal data APIs. In panel (b), the third variable is the standardization of identity and access management (IAM) software. Confidence intervals are shown in round brackets (instead of standard errors), estimated using wild cluster bootstrap with N=9999 which generates more conservative CI estimates than cluster robust standard errors when the data structure consists of a small number of large clusters. Columns 3 and 4 show regression results after applying propensity score matching (PSM) in column 3 and coarsened exact matching (CEM) in column 4 to pre-treatment (2017) variables log #PCs and cloud presence. *P*-values are shown in square brackets. $***p < 0.01; **p < 0.05; *p < 0.1.$

(a) Results on Intrafirm Data Interoperability

Dependent Variable	Log (Revenue)			
	(1)	(2)	(3) PSM	(4) CEM
TREAT × POST	-0.030** (-0.120, -0.003) [0.023]	-0.115** (-0.125, -0.091) [0.017]	-0.072** (-0.084, -0.043) [0.018]	-0.092** (-0.109, -0.059) [0.018]
API × POST		0.021 (-0.017, 0.104) 0.207	0.044* (-0.100, 0.135) 0.078	0.023* (-0.014, 0.100) 0.099
API × TREAT × POST		0.086* (-0.008, 0.135) [0.057]	0.063* (-0.019, 0.112) [0.073]	0.069* (-0.015, 0.120) [0.071]
Year FE	Y	Y	Y	Y
Establishment FE	Y	Y	Y	Y
Observations	86,555	86,555	61,330	86,550

(b) Results on Standardization of Identity and Access Management Software

Dependent Variable	Log (Revenue)			
	(1)	(2)	(3) PSM	(4) CEM
TREAT × POST	-0.024** (-0.146, -0.006) [0.003]	-0.188** (-0.355, -0.030) [0.029]	-0.229** (-0.442, -0.046) [0.028]	-0.178** (-0.332, -0.021) [0.032]
STD × POST		-1.450 (-3.688, 0.096) [0.115]	-3.501 (-22.577, 1.690) [0.104]	-1.649 (-4.961, 0.659) [0.095]
STD × TREAT × POST		4.134** (0.144, 8.231) [0.046]	6.186** (1.407, 11.487) [0.030]	4.137** (0.389, 7.890) [0.039]
Year FE	Y	Y	Y	Y
Establishment FE	Y	Y	Y	Y
Observations	65,740	65,740	43,690	65,740

Table 4: Effects of GDPR Enforcement on IT Investments and Job Postings by Data APIs. The tables show difference-in-differences regression results on establishment-year panel data, in subsamples split by the availability of data APIs. The outcome variables in panel (a) are different categories of IT budget in CITDB, and the outcome variables in panel (b) are defined based on the 2010 SOC occupation codes for IT occupations (15-11XX)**.** The tables show regression results on establishment-year panel data, in subsamples split by the availability of data APIs in panel (a) and median standardization of identity and access management (IAM) software in panel (b). Confidence intervals are shown in round brackets (instead of standard errors), estimated using wild cluster bootstrap with N=9999 which generates more conservative CI estimates than cluster robust standard errors when the data structure consists of a small number of large clusters. P-values are shown in square brackets. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

(a) IT Investments

Dependent Variable	Ln(IT Services Spend)		Ln(Hardware Spend)		Ln(Software Spend)	
Sample:	(1)	(2)	(3)	(4)	(5)	(6)
$I(API = 1)$	No	Yes	No	Yes	No	Yes
$TREAT \times POST$	-0.087**	0.002	-0.080***	0.002	-0.083**	0.002
	(-0.265, -0.026)	(-0.021, 0.017)	(-0.264, -0.017)	(-0.023, 0.017)	(-0.260, -0.023)	(-0.019, 0.016)
	[0.011]	[0.819]	[0.010]	[0.811]	[0.012]	[0.756]
Year FE	Y	Y	Y	Y	Y	Y
Estab. FE	Y	Y	Y	Y	Y	Y
Obs.	27,630	58,925	27,630	58,925	27,630	58,925
# Establishments	5,526	11,785	5,526	11,785	5,526	11,785
# Years	5	5	5	5	5	5

(a) IT Hiring

Dependent Variable	Ln(#IT Job Postings)		Ln(#IT-Info Job Postings)		Ln(#IT-Reg Job Postings)	
Sample:	(1)	(2)	(3)	(4)	(5)	(6)
$I(API = 1)$	No	Yes	No	Yes	No	Yes
$TREAT \times POST$	0.795***	0.134	0.521**	0.097	0.701***	0.159
	(0.427, 0.959)	(-0.204, 0.445)	(0.439, 0.643)	(-0.140, 0.349)	(0.491, 0.861)	(-0.142, 0.461)
	[0.009]	[0.386]	[0.011]	[0.422]	[0.009]	[0.252]
Year FE	Y	Y	Y	Y	Y	Y
Firm-State FE	Y	Y	Y	Y	Y	Y
Obs.	1,770	3,870	1,770	3,870	1,770	3,870
# Firm-State	118	258	118	258	118	258
# Quarters	15	15	15	15	15	15

Table 5: Effects of GDPR Enforcement on IT Investments and Job Postings by Standardization. The tables show difference-in-differences regression results on establishment-year panel data, in subsamples split by the median standardization of identity and access management (IAM) software. The outcome variables in panel (a) are different categories of IT budget in CITDB, and the outcome variables in panel (b) are defined based on the 2010 SOC occupation codes for IT occupations (15-11XX). The last 2-digits of the SOC code are 11, 21, or 22 for Computer and Information Research Scientists & Analysts. The last 2-digits of the SOC code are 41, 42, or 43 for Database & Systems Admin & Network Architects. The last 2-digits of the SOC code are 69, 70, 71, or 72 for Software Developers & Programmers. Confidence intervals are shown in round brackets (instead of standard errors), estimated using wild cluster bootstrap with N=9999 which generates more conservative CI estimates than cluster robust standard errors when the data structure consists of a small number of large clusters. *P*-values are shown in square brackets. $***p <$ *0.01;* $**p <$ *0.05;* $*p <$ *0.1.*

(a) IT Investments

Dependent Variable	Ln(IT Services Budget)		Ln(Hardware Budget)		Ln(Software Budget)	
Sample:	(1)	(2)	(3)	(4)	(5)	(6)
$I(STD \geq P50)$	No	Yes	No	Yes	No	Yes
TREAT × *POST*	-0.039**	-0.002	-0.035*	-0.000	-0.035*	-0.005
	(-0.163, 0.000)	(-0.344, 0.546)	(-0.160, 0.002)	(-0.355, 0.630)	(-0.154, 0.002)	(-0.346, 0.573)
	[0.049]	[0.444]	[0.072]	[0.823]	[0.074]	[0.157]
Year FE	Y	Y	Y	Y	Y	Y
Estab. FE	Y	Y	Y	Y	Y	Y
Obs.	27,430	38,310	27,430	38,310	27,430	38,310
# Establishments	5,486	7,662	5,486	7,662	5,486	7,662
# Years	5	5	5	5	5	5

(b) IT Hiring

Dependent Variable	Ln(#IT Job Postings)		Ln(#IT-Info Job Postings)		Ln(#IT-Reg Job Postings)	
Sample:	(1)	(2)	(3)	(4)	(5)	(6)
$I(STD \geq P50)$	No	Yes	No	Yes	No	Yes
TREAT × *POST*	0.145	0.821	0.149	0.504	0.192	0.666
	(-0.358, 0.887)	(-0.615, 1.558)	(-0.119, 0.639)	(-0.481, 2.060)	(-0.227, 0.831)	(-0.667, 1.444)
	[0.586]	[0.187]	[0.311]	[0.292]	[0.390]	[0.246]
Year FE	Y	Y	Y	Y	Y	Y
Firm-State FE	Y	Y	Y	Y	Y	Y
Obs.	1,590	1,695	1,590	1,695	1,590	1,695
# Firm-State	106	113	106	113	106	113
# Quarters	15	15	15	15	15	15

Online Appendix – Data Governance, Interoperability and Standardization: Organizational Adaptation to Privacy Regulation

Table of Contents

List of Appendix Figures

List of Appendix Tables

Appendix A1. Data Construction.

A1 Data Sources.

The paper's primary empirical analyses rely on merging information from three data sources at the establishment level and at the corporation level. We describe these data sources as follows.

A1.1 Aberdeen Computer Intelligence Technology Database (CITDB).

The CITDB is the most authoritative data source on establishment-level IT products and widely used in prior research in information systems, management, finance, and economics (e.g., Bresnahan, Brynjolfsson, & Hitt, 2002; Kretschmer, 2004; Forman, 2005; Forman, Goldfarb, & Greenstein, 2005; Dewan, Shi, & Gurbaxani, 2007; Chwelos, Ramirez, Kraemer, & Melville, 2010; Bloom, Sadun, & Van Reenen, 2012; Kretschmer, Miravete, & Pernías, 2012; McElheran, 2014; Haug, Kretschmer, & Strobel, 2016; Nagle, 2019; Tuzel & Zhang, 2021; Huang, Ceccagnoli, Forman, & Wu, 2022; Sambhara, Rai, & Xu, 2022). We use only the U.S. (not European) data snapshots because the setting of this paper focuses on large multinational corporations based in the United States. The database contains annual snapshots with similar data structures. We combine five years of data on U.S.-based establishments from 2016 to 2020.

However, this data source suffers a few data quality issues, which we describe as follows.

1. The number of firms with non-missing IT product data vary from year to year, and in some years, the sample contains more establishments than others;
2. Not all the variables were consistently measured, and many variables were imputed prior to 2016;
3. There is a major change in data collection methodology in 2016—2017 such that (1) the number of sampled establishments increased drastically with a much-improved coverage rate and (2) comparing the 2015 and 2016 snapshot suggest that many variables suffer discrepancy between these years, however not between consecutive years since 2016.
4. In 2016, the revenue variable contained largely zeros (95% of the data relative to less than 10% in subsequent years), which we believe are missing values after conducting substantial checks

In addition, the 2020 snapshot contains the most detailed information. For example, the revenue variable (in $millions) was rounded to an integer number before 2020, but the exact value was reported in 2020. The IT products dataset contains substantially more comprehensive coverage of vendors and products and added a new variable that records when a product was first seen. We discuss the measurement of establishment annual revenues in more detail in Appendix Section A3. We describe all the main variables used in the empirical analysis in Appendix Section A5.

A1.2 Keystone-Microsoft "Tech Intensity Scorecard" Survey.

In 2020, Keystone Strategy LLC and Microsoft Corporation jointly developed and conducted a detailed survey measuring large corporations' AI technology intensity. The original survey contains over a hundred questions developed based on deep industry expert knowledge about bleeding-edge infrastructural and software technologies around data, cloud, and machine learning capabilities. The survey was designed to be a standardized framework for assessing the maturity of enterprise data architecture in large corporations across data-intensive industries with digital and traditional operating models. Questions focus on a detailed set of technical frameworks, objects, and capabilities rather than broad and high-level strategies. The questions were designed to elicit yes/no responses about objective information, which makes it easy to code up the data and define clear-cut categories split on a single variable.

The survey was administered to senior technology executives, e.g., chief information officers (CIOs) and chief digital officers (CDOs) in charge of corporation-level IT initiatives. The corporations are all very large U.S.-based conglomerates. The sample consists of Fortune 500 corporations, and the survey team reached out to executives through LinkedIn and other connections. While each corporation spans hundreds of establishments, the survey questions are about organization-wide (hence corporation-level) capabilities rather than those within a single business unit. The data collection team conducted in-person interviews to collect the survey responses.

This paper's empirical analyses use the survey data on one question: *"Do you use services interfaces or APIs to publish departmental (e.g., not application) information between departments and/or functional areas?"*. Answers to this question are binary, and Appendix Table C1 shows the raw data containing each corporation's response.

A1.3 Corporations' Annual 10-K Filings.

The United States Securities and Exchange Commission (SEC) requires public companies to file annual reports (Form 10-K), which disclose crucial information such as the firm's operating results, financial conditions, managerial approaches, and relevant regulations. These 10-K filings are publicly available, and we download the raw files in PDF format directly from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) service.

All of the corporations in our sample are public companies, for which we collect annual reports (10-K Forms) from 2014 to 2017. We manually extract three types of information from these annual reports.
First, we collect information about the geographic distribution of revenues, credit risks, and business activities from the financial statements and textual descriptions. Second, we collect the keywords mentioning Europe and European countries. Third, we can also observe texts discussing the GDPR and the corporation's compliance efforts.

Our primary purpose for using the 10-K filings data is to understand the extent to which the GDPR affects each corporation differentially due to the variation in their exposure to the regulation, depending on the share of business activities involving consumers in Europe. Appendix Section A4 contains a detailed description of how we use the 10-K filings data to measure each corporation's exposure to the European consumer market. Appendix Figures B2 and Table C3 contain supporting information summarized from the 10-Ks.

A2 Matching Establishments to Corporations.

Our data sources are at two levels: the CITDB data is at the establishment, and the 10-K filings and survey data are at the corporation level. Hence, matching these data sources requires identifying the establishments in CITDB that belong to each corporation in the sample. This is a non-trivial task that requires significant effort. The data provider has some information about the corporation associated with each establishment. However, the information is not always accurate. Hence, careful data cleaning is required to ensure that establishments included for data analyses indeed belong to the matched corporation in the final sample. The major challenge to creating a clean mapping between establishments and a corporation is that large corporations frequently undergo numerous events every year that change the ownership status of establishments. We do not have access to any existing establishment-level data source that keeps track of all these changes in a systematic manner and can be matched to the CITDB data directly. Furthermore, the raw data contain plain mismatches that require correction, possibly because the data provider used automated text-based methods (e.g., machine learning and natural language processing) to match all the establishments. For more details, Crane & Decker (2019) contains a useful account of challenges associated with identifying establishments of large U.S. companies more generally.

The organizational changes that we need to check for mostly fall under the following categories:

- Merger and acquisition: An establishment changes ownership and hence may belong to a corporation this year but another corporation in the next year
- Spin-offs and liquidation: An establishment is sold or liquidated by the corporation to another company or a private equity firm, or the establishment closes down and must be removed from the data
- Name change: A sub-unit of the corporation re-brands itself and changes its name

We match establishments to a corporation using a two-step procedure. The first step is an automated approach, and the second is a manual one.

First, we start with an automated approach designed to identify *all* the establishments (or as exhaustively as possible) that may belong to the corporation. In this step, we emphasize avoiding false negatives with more tolerance for false positives, which can be dealt with later in the manual step. The automated approach consists of three sub-steps as follows:

1. Determine each corporation's potential U.S. headquarter(s): In this step, we find all the possible candidate establishments in CITDB that can be the corporation's U.S. headquarter. To do so, we look up the DUNS number of the corporation's headquarter in Mergent Intellect (a business intelligence platform that aggregates company profiles) and match them to the establishment in the CITDB with the same DUNS number. For each year from 2016 to 2019, we additionally perform fuzzy matching on the name and address of the U.S. headquarter reported in Compustat and then select the establishment in the CITDB that is the closest match.
2. Find all the establishments that are likely subsidiaries of each corporation: For each annual snapshot of the CITDB, we find all the establishments with the same EnterpriseID or URL domain as the headquarter(s) identified in sub-step 1. We do this for each year's data separately.
3. Include all years of CITDB data for each establishment: We match each establishment found in sub-step 2 to all years of CITDB data. This includes information in all other annual snapshots about the same establishment found in one year. This step is needed because some establishments appear to have different EnterpriseIDs linked to them in CITDB in different years, which can be due to either an actual change in ownership or a mismatch in the raw data in one or more years. Since we want to be as exhaustive as possible in the automated step, we include all the observations associated with establishments that are potentially part of the sampled corporations in at least one year between 2016 and 2020.

Second, we conduct manual searches to rule out erroneous matches. The CITDB data is not always accurate in linking an establishment to the correct EnterpriseID. We perform manual data cleaning to weed out establishments mistakenly linked to a corporation. We manually search for the name of each establishment obtained in the previous automated step and only include them if they are in the list of subsidiaries of the corporation during the focal year. For example, we remove establishments spun off (or split) from the corporation before the focal year from the data. Establishments that became part of the corporation due to a merger or acquisition may not be part of it in an earlier year and should thus be removed from the data in the earlier year. More generally, if an establishment changed ownership during the sample period, we drop it from the data to ensure that the sample consists of a stable set of establishments that are not at risk of experiencing significantly different events that may interfere with identifying the impact of GDPR.

We then review the names of each of the 25 financial services corporations' list of establishments to weed out those erroneously classified as belonging to a corporation in the sample. To ensure that an establishment identified as a subsidiary of a corporation in each year indeed operated as such in that year, we go through the name of each establishment that results from the previous (automated) step to check manually and delete it from the data if it was misclassified as a subsidiary of the corporation. The 10-K reports of the corporations contain lists of subsidiaries each year, and we check whether the establishment's name appears in these lists. Establishments are dropped if they fall under one of the following three cases:

- Completely mistaken match: the name (of the subsidiary) cannot be found at all in association with the corporation
- The establishment became part of the corporation as a result of a merger or acquisition that occurred *after* 2016; that is, the establishment must already be part of the corporation when the sample period begins
- The establishment was spun off or split from the corporation *in or after* 2016; that is, the establishment must be part of the corporation in all the observed snapshots during the sample period

Finally, the establishment must satisfy the following two conditions to be included in the final sample.

1. Establishments operating businesses unrelated to personal data, namely those with NAICS codes indicating monetary authorities (521XXX) and investment banking (523110), are excluded from the sample.
2. To ensure that the sample is a balanced panel of establishments, we include only establishments with non-missing data for all years of the sample period (2016–2020); we impose this requirement after the imputation of revenue data described in Appendix Section A3.

A3 CITDB Data Issues and Imputing Revenues in 2016

The data collection methodology of the firm providing the data requires further elaboration to provide more clarity into the quality of this data and especially to inform future researchers who wish to combine multiple years of the CITDB into a panel dataset to study changes in establishment-level variables across years.

Plotting aggregate trends in the raw data led us to believe there was a sharp change in how the data was collected before and after 2016. This confirms the finding of recent empirical research using the CITDB establishment data (e.g., Chen, Balasubramanian, & Forman, 2022). The historical CITDB data was collected through survey interviews. However, the data collection changed around 2016–2017 from a survey to a richer set of information sources and web-scraping to obtain online information, including the resumes of business users indicating experience or involvement with different technologies (e.g., Levy, 2019). The change in data collection methodology by the CITDB data provider makes comparing data across years before and after 2016 untenable. In addition, the CITDB data may suffer declining quality after its ownership status change in 2006 and before it was split from Harte Hanks around 2015 (Levy, 2015).

The data has a broader coverage of establishments with available technology product data since 2017 than before in the CITDB. From 2015 to 2016, there appears to be a sharp change in the revenue variable across all establishments in the data. Establishments are identified by a unique SiteID across multiple years, which allows us to link observations on the sample establishment across yearly snapshots of the CITDB data. We check the consistency of the measurement of the revenue variable across years by calculating the autocorrelation across one year and the year before based on the entire CITDB database of U.S. establishments. In 2016, the correlation between log establishment revenue and lagged log revenue was 0.5637. In 2017, the correlation between log establishment revenue and lagged log revenue was 0.9683. In 2016 and 2017, the autocorrelation statistics were calculated based on establishments reporting non-zero revenues in both years, which account for about 37% of all the observations. In 2018, the correlation between log establishment revenue and lagged log revenue was 0.9765. In 2019, the correlation between log establishment revenue and lagged log revenue was 0.9758. In 2020, the correlation between log establishment revenue and lagged log revenue was 0.8975. From 2018 to 2020, the autocorrelation statistics were calculated based on establishments reporting non-zero revenues in both years, which accounted for about 83% of the entire database. In all years before 2020, the CITDB reports revenues rounded to the nearest integer in $ millions. In 2020, the CITDB reports revenues that are not rounded. This difference in rounding explains the somewhat lower autocorrelation in 2020.

Since the revenue data are available for all publicly traded U.S. firms through Compustat, we can compare the autocorrelation in the actual revenues (at the corporation level) with the statistics derived from CITDB. Between and 2020, the autocorrelation of log company revenue among all Compustat firms is 0.9786. This is much closer to the autocorrelation statistics in the CITDB data from 2017 to 2020, but not between 2015 and 2016, which is too low. In addition, in the sample of corporations we study, only 5% of the establishments in 2016 reported non-zero revenues, in contrast to over 90% in 2017 and every year after. It is highly plausible that most of the zeros in the 2016 revenue data are missing values, consistent with the data provider making structural changes to the data source after Aberdeen (the underlying company) split from Harte Hanks and began to upgrade its data collection methodology around mid-2015, which may take some time to reflect in the eventual data product that it offers to researchers. This also means that the CITDB data is inconsistent when the year crosses from 2015 to 2017. Hence, the database should not be used to construct panel data containing information both before and after 2016.

Next, we address the missing revenue data problem in the 2016 snapshot of CITDB. The fact that a substantial fraction of revenues in 2016 are missing and that the data prior to 2016 are not usable pose challenges to econometric identification. The number of time periods with a sufficiently large share of non-missing observations before treatment (i.e., GDPR enforcement in 2018 for treated firms) is too small. To assess the parallel trends assumption (PTA), we need at least two time periods before 2018 with enough non-missing observations.

We address this issue by filling the missing 2016 revenue data with imputed values following the procedure described below. To do this, we need data on corporation-level annual revenues in both 2016 and 2017 (which can be accessed from corporations' public 10-K filings) and data on establishment-level revenues in CITDB in 2017. We can rewrite the establishment-level revenue in 2016 using the following accounting identity.

$$Revenue_{c,i,2016} = CorpRev_{c,2016} \times \frac{Revenue_{c,i,2016}}{CorpRev_{c,2016}} = CorpRev_{c,2016} \times RevShare_{c,i,2016}$$

We then calculate the imputed establishment-level revenues in 2016 as follows. We impute the revenue of establishment i in corporation c by replacing the unknown $RevShare_{c,i,2016}$ with known $RevShare_{c,i,2017}$, which equals the establishment revenue divided by corporation total revenue in 2017. Hence

$$ImputedRev_{c,i,2016} = CorpRev_{c,2016} \times RevShare_{c,i,2017} = CorpRev_{c,2016} \times \frac{Revenue_{c,i,2017}}{CorpRev_{c,2017}}$$

If exposure to the European consumer market is associated with differential trends in pre-2018 revenue, we should be able to detect such trends using the imputed establishment revenue. The imputed revenue is calculated based on the corporation's total revenue multiplied by the approximate share. It should thus reflect meaningful changes in revenue between 2016 and 2017 because it is calculated based on actual observed corporation revenue data.

Moreover, none of the corporations in the sample experienced major re-organization or merger events in 2016 – 2017. Hence, the establishments did not experience critical changes in their share of total corporation revenues. Econometric identification thus relies on the difference in imputed and actual revenues being an exogenous random variable independent of treatment status. If firms with high exposure to the European market experience systematically different revenue growth from 2016 to 2017 relative to other firms, then the total corporation-level revenue growth should reflect this difference, and the pre-trend should be identified in the imputed establishment data as well.

To illustrate how well the imputation works, Appendix Figure B1 panel (a) compares the distribution of establishment revenues in the CITDB 2016 between actual and imputed data for those with non-missing data in 2016. The distributions are very similar, suggesting that the imputed revenues track the actual revenues (and the correlation is 0.9685). Table C2 shows descriptive statistics of the actual and imputed revenues for the subsample with non-missing values and the full sample in 2016. Appendix Figure B1 panel (b) compares the distributions of the imputed revenues for all establishments with the distribution of actual revenues for those with non-missing values. The imputed values have a lower average and distribution skewed to the left. This suggests that the missing 2016 revenues are associated with smaller establishments with lower revenues.

A4 Measuring the Exposure to GDPR Compliance Risks

Each corporation's exposure to GDPR-related risks is measured by the share of revenues within the GDPR jurisdiction (Europe) affected by the regulation. We collect relevant information from corporations' 10-K filings in 2014 and 2015 (before the GDPR announcement) to measure each corporation's exposure to the European consumer market. The final exposure measure is the average of the measures in 2014 and 2015, although the values are very similar between these two consecutive years. We describe our methodology as follows.

Scenario data source [1] in Table 1: We begin by looking for information on the geographic distribution of revenues, especially the share of revenues from European operations. The only exception is revenue from investment banking, which should be excluded in calculating exposure to the European consumer market because it does not involve personal data as the clients are primarily institutional (non-personal) entities, including very large companies. This applies to only the following corporations in the sample: JPMorgan Chase, Goldman Sachs, and Bank of America. Also, many corporations lump together revenues from Europe, the Middle East, and Africa (EMEA) and do not separately report a revenue number for Europe only. However, in most of these cases, the European market dominates the EMEA revenues. Hence, we can use the EMEA revenue to calculate exposure to the European consumer market. Using this approach, we define the exposure measure for many of the corporations in the sample, including BNY Mellon, Goldman Sachs, Citigroup, Morgan Stanley, American Express Company, American International Group, Bank of America Corp, MetLife, JPMorgan Chase, Capital One, The Travelers Companies, Lincoln National Corporation, and Centene Corporation. The exact variable equals the European (or EMEA) revenues from business segments involving personal data divided by total corporation revenue.

Scenario data source [2] in Table 1: For corporations without sufficient revenue information to determine the share of revenues accounted for by the European consumer market, we use the total credit risk exposure to European countries as a fraction of total interest-earning assets. The 10-K filings report the total credit risk of European exposure either directly or separately by country (in which case, we sum up the values across all European countries listed in the report). Using this approach, we define the exposure measure for Markel Corporation, Assurant, Hartford Financial Services Group, Fifth Third Bancorp, Wells Fargo, and Charles Schwab.

Scenario data source [3] in Table 1: For the remaining corporations, a direct breakdown of neither revenues nor credit risk by geography is available for Europe or the EMEA region. Insurance companies' primary business activity involves underwriting to determine clients' risks. Many insurance companies report the geographic distribution of net premiums earned. Hence, we define the exposure measure as the share of net premiums earned in Europe (or the EMEA region) if this information is available. This applies to one insurance company, specifically Chubb Limited.

Scenario data source [4a] and [4b] in Table 1: Finally, if none of the above is available, we derive the exposure measure using the share of international (outside the U.S.) revenues. However, international revenues often come from regions such as Canada, Australia, and countries in Asia and Africa. To determine whether the non-US revenue primarily reflects business activities in Europe or non-European countries, we read through the 10-K filings in more detail to find qualitative evidence on whether Europe is the corporation’s primary foreign market. Among these corporations, Europe is the primary foreign market for MoneyGram International, Mastercard, and Eaton Vance, while it is not the primary foreign market for Prudential Financial and UnitedHealth Group. If Europe is the primary foreign market, then we define the exposure measure as non-US revenue from business segments involving personal data divided by total corporation revenue [4a]. If Europe is not the primary foreign market, then the exposure measure equals zero [4b]. Appendix Table C3 summarizes textual information in the 10-K filings about these corporations’ business operations worldwide, which are used to determine whether Europe is the primary foreign market. The revenues from Europe are trivial in the latter case [4b] when the primary foreign markets are other countries that account for the corporations’ primary revenue activities outside the U.S.

The primary data-intensive business segments in which the corporations in the sample operate are payment, consumer banking, insurance, and asset management. GDPR compliance is very important to corporations with substantial European consumer business activities. Most sample corporations discuss GDPR in the Relevant Regulations section in their 10-K filings and mention how they implement new approaches to comply with the regulation. The only corporations not mentioning the GDPR in 2017 or before are Eaton Vance Corp, Fifth Third Bancorp, Wells Fargo, Centene Corporation, Lincoln National Corporation, and Charles Schwab. The exposure to the European consumer market is 5% or below for all of these corporations, according to the measurement approach described in this section.

Finally, Appendix Figure B2 plots the measured exposure to the European consumer market against the number of times a keyword such as Europe or a major European country (i.e., United Kingdom, Ireland, Germany, France, Italy, or Spain) appears in the 10-K text in 2015. There is an evident positive correlation between the frequency of European keywords and our derived measure of exposure to the European consumer market, which supports the latter's validity.

A5 Variable Descriptions

Annual Establishment Revenue. This is the establishment performance measure in annual Aberdeen CI Technology Database snapshots from 2016 to 2020. We take the natural logarithm of the revenues measured in U.S. dollars to be the outcome variable. See Appendix Sections A1 and A3 for details.

GDPR Exposure. This is the corporation's exposure to GDPR compliance risks, using data from 10-K filings in 2014 and 2015 (prior to the announcement of the GDPR). The exposure measure is defined by the corporation's share of revenues, credit risks, or business activities in the European consumer market. The variable ranges from 0 to 1. See Appendix Sections A4 for details.

Internal Data APIs. This is the binary indicator of whether the corporation uses services interfaces (APIs) to publish data or information across departments and functional areas. The measurement of this variable is based on survey data from the Keystone-Microsoft collaboration. See Appendix Section A1 for details.

Standardization of IAM Software. This measures the standardization of identity and access management (IAM) software vendors at the establishment level in 2017. This variable is constructed based on the presence of different IAM software products. See main paper Section 3.3 for details.

#PCs. This measures the number of personal computers (PCs) at the establishment level in 2017.

Cloud Presence. This measures the presence of one or more of the following public cloud providers – Amazon, Microsoft, Google, or IBM in 2017.

ΔRevenue (2016-17). This measures the change in the log establishment revenue from 2016 to 2017.

IT Services Spend. Establishment level annual spending on IT services (in US$).

Hardware Spend. Establishment level annual spending on computing hardware (in US$).

Software Spend. Establishment level annual spending on computing software (in US$).

#IT Job Postings. Firm-state level quarterly number of job postings for Computer Occupations Except Support Specialists (2010 SOC codes all of 15-11XX except 15-115X).

#IT-Info Job Postings. Firm-state level quarterly number of job postings for Computer and Information Research Scientists & Analysts (2010 SOC codes 15-111X and 15-112X).

#IT-Reg Job Postings. Firm-state level quarterly number of job postings for Computer Occupations with Regulation Related Tasks (2010 SOC codes 15-1111, 15-1121, 15-1134, 15-1141, and 15-1199). The included SOC codes have at least one regulation-related task in Trebbi, Zhang, & Simkovic (2023).

A6 Corporation-Level Analyses for Subsamples Split by Internal Data APIs

We conduct robustness checks using aggregated data obtained by averaging the variables across all establishments within the same corporation in each year from 2016 to 2020. In the original panel data, we excluded establishments that either became part of or split from the corporation during the sample period to obtain a balanced panel. Hence, each observation in the aggregated data for a corporation consists of data from the same set of establishments over time. That is, the year-to-year change in the observations in the aggregate data reflects solely revenue changes within the same establishments, but not due to new establishments being added or dropping out of the data. We estimate the same model in Equation 2 on the aggregated data and control for corporation fixed effects and year fixed effects.

Appendix Figures B4 plot the estimates from the event-study version of the regression model. In contrast to many prior studies that observe abrupt but temporary changes immediately after GDPR enforcement, which become smaller gradually over time, we observe a gradual decline in revenue that becomes more significant over time among corporations that lack internal data APIs. The revenue changes we find are likely due to both supply and demand factors. They reflect firms' compliance efforts through removing or installing GDPR-compliant technologies and changes in consumers' perception of the quality of products and services provided by the firm. The fact that revenue declines slowly rather than abruptly among organizations for which balancing compliance with business continuity is more challenging is not surprising. The organizations studied in our setting are all very large legacy firms, for which making system-wide adjustments is much more difficult and takes longer due to organizational inertia. Many of these organizations already struggled with combining heterogeneous data silos and taming convoluted IT systems, which were much stickier than those in startups and smaller firms.

Appendix Table C6 reports the regression results for the sample splits. A primary concern for using the event-study design to estimate the effects of GDPR enforcement is that corporations may have responded to the regulation by making efforts to comply before the enforcement date. The financial services industry is already subject to regulations regarding information sharing under existing U.S. laws such as the Gramm-Leach-Bliley Act (GLB) and the Health Insurance Portability and Accountability Act (HIPAA). These laws apply to the corporations in the sample as well. Relative to the average-sized firm, the large mature organizations in the sample are both subject to more significant penalties if they violate the GDPR and are more likely to have historically developed regulatory risk management practices. To address the potential violation of the parallel trends assumption (PTA), we estimate the regression results using the synthetic difference-in-differences (SDID) method (Arkhangelsky et al., 2021). We use the average number of PCs and the share of establishments with the cloud as the covariates for minimizing the distance between synthetic counterfactuals and treated units. The SDID method mitigates the non-parallel trends problem by constructing synthetic control counterfactuals close to the treated units.

References

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

Bresnahan, T., Brynjolfsson, E., & Hitt, L. M. (2002). Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence. *The Quarterly Journal of Economics*, *117*(1), 339–376.

Chen, R., Balasubramanian, N., & Forman, C. (2022). *How Does Labor Mobility Affect Business Adoption of a GPT? The Case of Machine Learning.*

Chwelos, P., Ramirez, R., Kraemer, K. L., & Melville, N. P. (2010). Research Note—Does Technological Progress Alter the Nature of Information Technology as a Production Input? New Evidence and New Results. *Information Systems Research*, *21*(2), 392–408.

Crane, L. D., & Decker, R. (2019). *Business Dynamics in the National Establishment Time Series (NETS)*.

Dewan, S., Shi, C., & Gurbaxani, V. (2007). Investigating the Risk–Return Relationship of Information Technology Investment: Firm-Level Empirical Analysis. *Management Science*, *53*(12), 1829–1842.

Forman, C. (2005). The Corporate Digital Divide: Determinants of Internet Adoption. *Management Science*, *51*(4), 641–654.

Forman, C., Goldfarb, A., & Greenstein, S. (2005). How Did Location Affect Adoption of the Commercial Internet? Global Village vs. Urban Leadership. *Journal of Urban Economics*, *58*(3), 389–420.

Haug, K. C., Kretschmer, T., & Strobel, T. (2016). Cloud Adaptiveness Within Industry Sectors–Measurement and Observations. *Telecommunications Policy*, *40*(4), 291–306.

Huang, P., Ceccagnoli, M., Forman, C., & Wu, D. J. (2022). IT Knowledge Spillovers, Absorptive Capacity, and Productivity: Evidence from Enterprise Software. *Information Systems Research*, *33*(3), 908–934.

Kretschmer, T. (2004). Upgrading and Niche Usage of PC Operating Systems. *International Journal of Industrial Organization*, *22*(8–9), 1155–1182.

Kretschmer, T., Miravete, E. J., & Pernías, J. C. (2012). Competitive Pressure and the Adoption of Complementary Innovations. *American Economic Review*, *102*(4), 1540–1570.

Levy, M. (2015). *Freed from Harte-Hanks, the AccessCI Database May Become Relevant Again.* https://gzconsulting.org/2015/10/03/freed-from-harte-hanks-the-accessci-database-may-become-relevant-again/

Levy, M. (2019). *Aberdeen Behavioral Technographics.* https://gzconsulting.org/2019/10/07/aberdeen-behavioral-technographics/

McElheran, K. (2014). Delegation in Multi-Establishment Firms: Adaptation vs. Coordination in I.T. Purchasing Authority. *Journal of Economics and Management Strategy*, *2*, 225–257.

Nagle, F. (2019). Open Source Software and Firm Productivity. *Management Science*, *65*(3), 1191–1215.

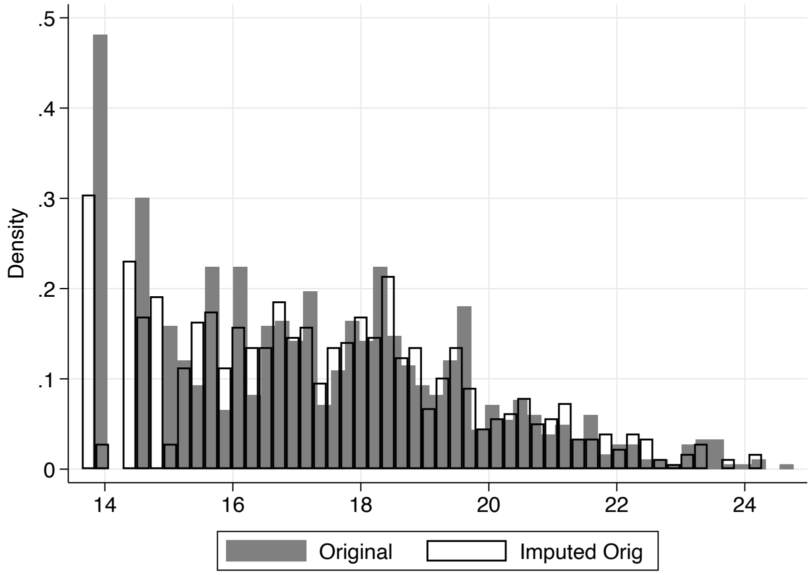
Sambhara, C., Rai, A., & Xu, S. X. (2022). Configuring the Enterprise Systems Portfolio: The Role of Information Risk. *Information Systems Research*, *33*(2), 446–463.

Trebbi, F., Zhang, M. B., & Simkovic, M. (2023). *The Cost of Regulatory Compliance in the United States.*

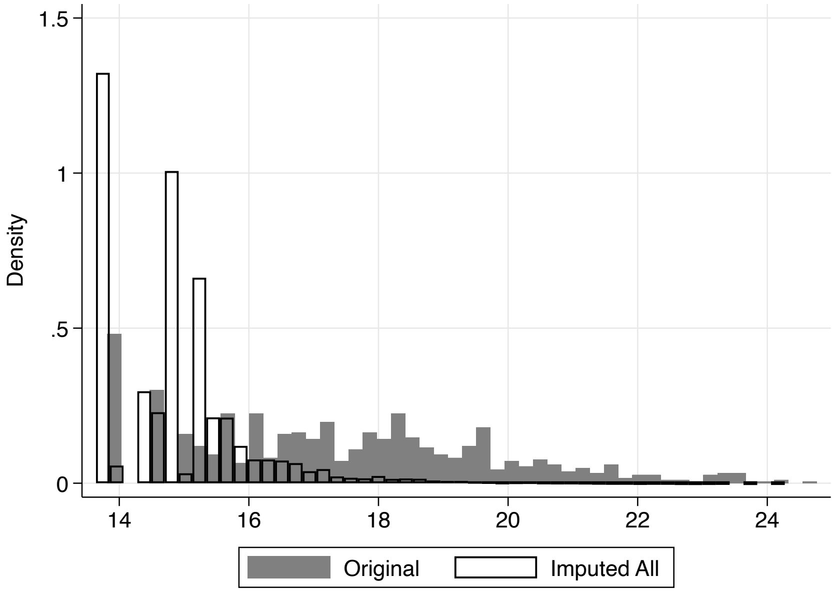
Tuzel, S., & Zhang, M. B. (2021). Economic Stimulus at the Expense of Routine-Task Jobs. *Journal of Finance*, *76*(6), 3347–3399.

Appendix Figure B1: Distribution of Establishment Revenues in 2016. Panel (a) compares the distribution of original and imputed log revenues among establishments with non-missing revenues in the CITDB snapshot in 2016 (N=866). Panel (b) shows the distribution of log revenues among establishments with non-missing revenues (N=866) and the distribution of imputed log revenues for all establishments in 2016 (N=36,500).

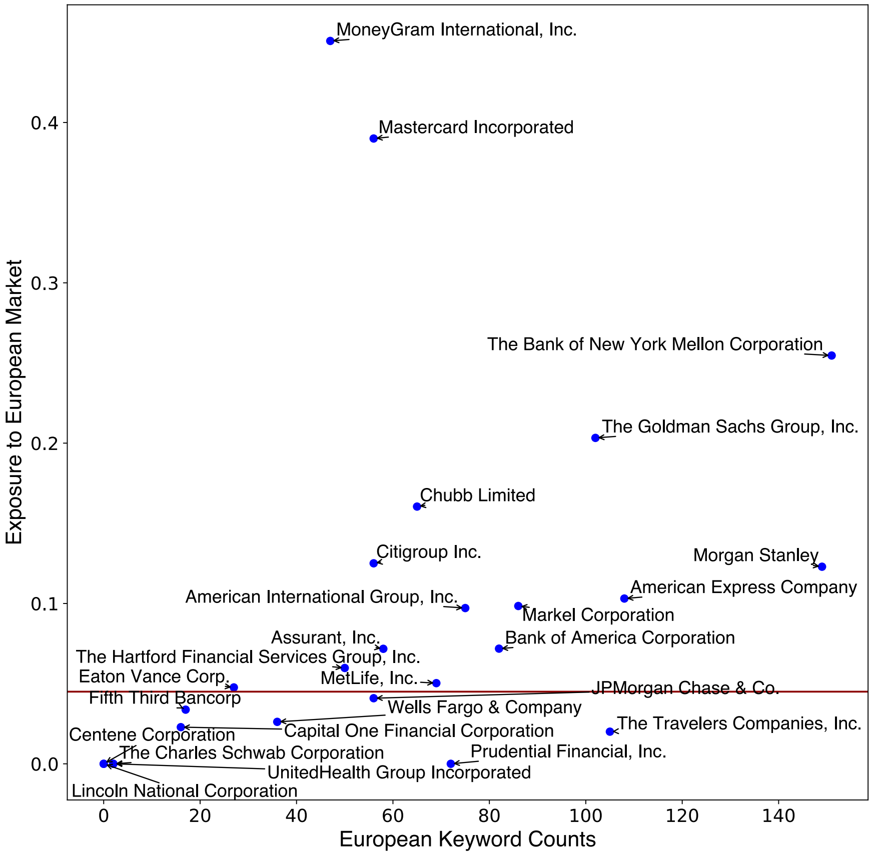
(a) Establishments with Non-Missing Revenues in 2016: Reported Revenue vs. Imputed Revenue



(b) Reported Non-Missing Revenues (N=866) vs. Imputed Revenues for All Establishments

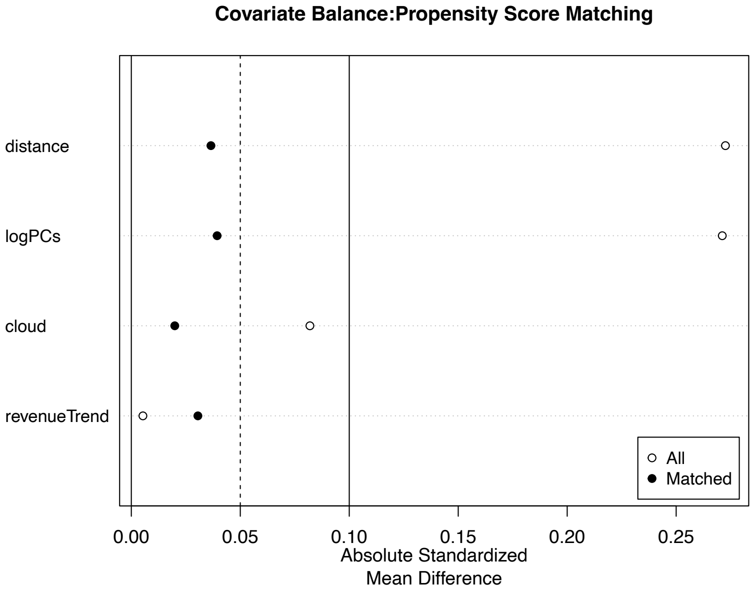


Appendix Figure B2: Corporations' Exposure to European Consumer Market. This figure plots each corporation's exposure to the European consumer market against the number of times either "Europe" or a major European country is mentioned in the 10-K. Both variables are measured based on information in the corporation's 10-K in 2015 (before the GDPR announcement). The correlation between the two variables is 0.3593. The horizontal line shows the threshold for defining the treatment indicator (5% rounded or y=0.045).

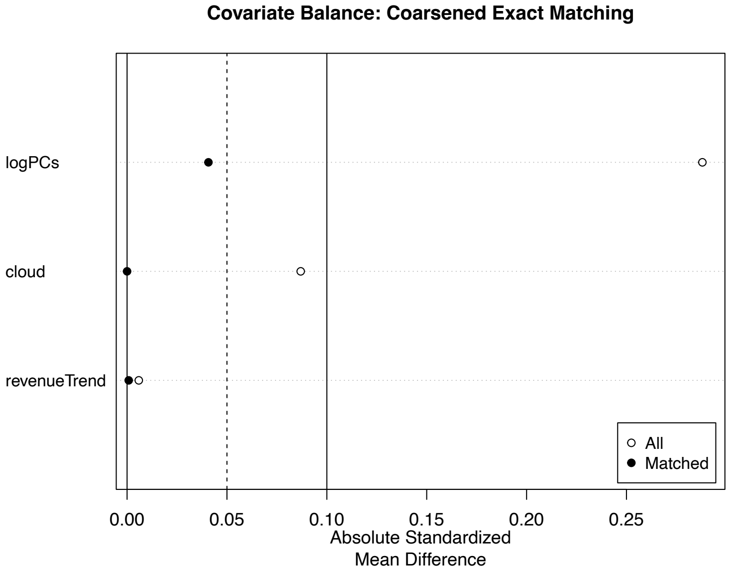


Appendix Figure B3: Covariate Balance Comparisons Before and After Matching. The covariate balance plots are based on the full panel of 17,311 establishments. The matching is conducted on log #PCs and cloud presence, measured in 2017. Panel (a) and (b) compare covariate balances before and after propensity score matching (PSM) and coarsened exact matching (CEM), respectively.

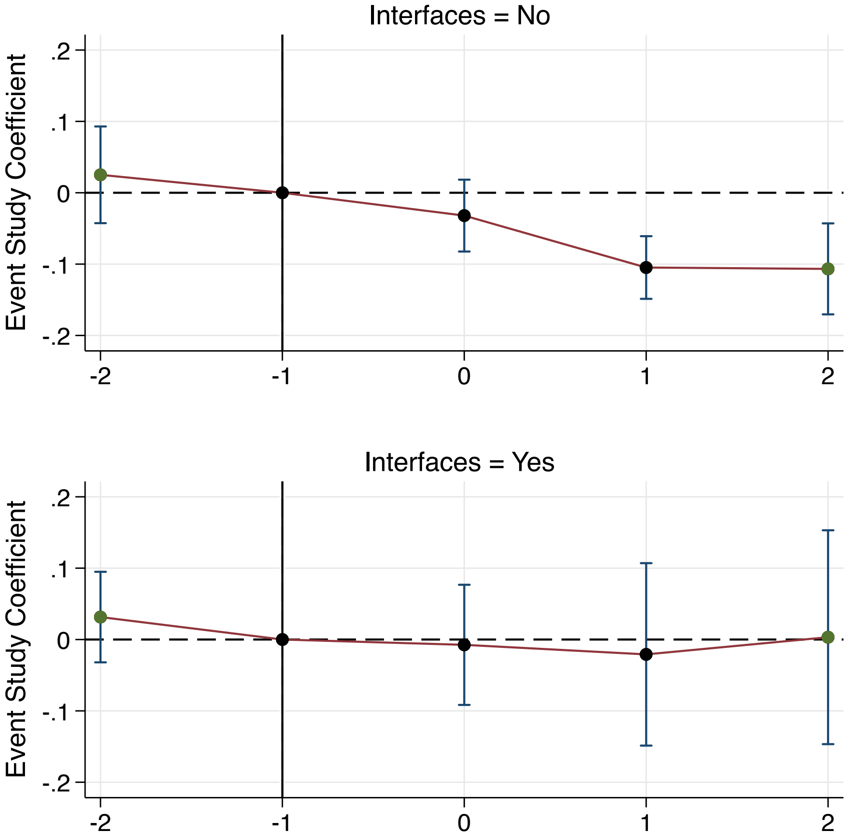
(a) Propensity Score Matching (PSM)



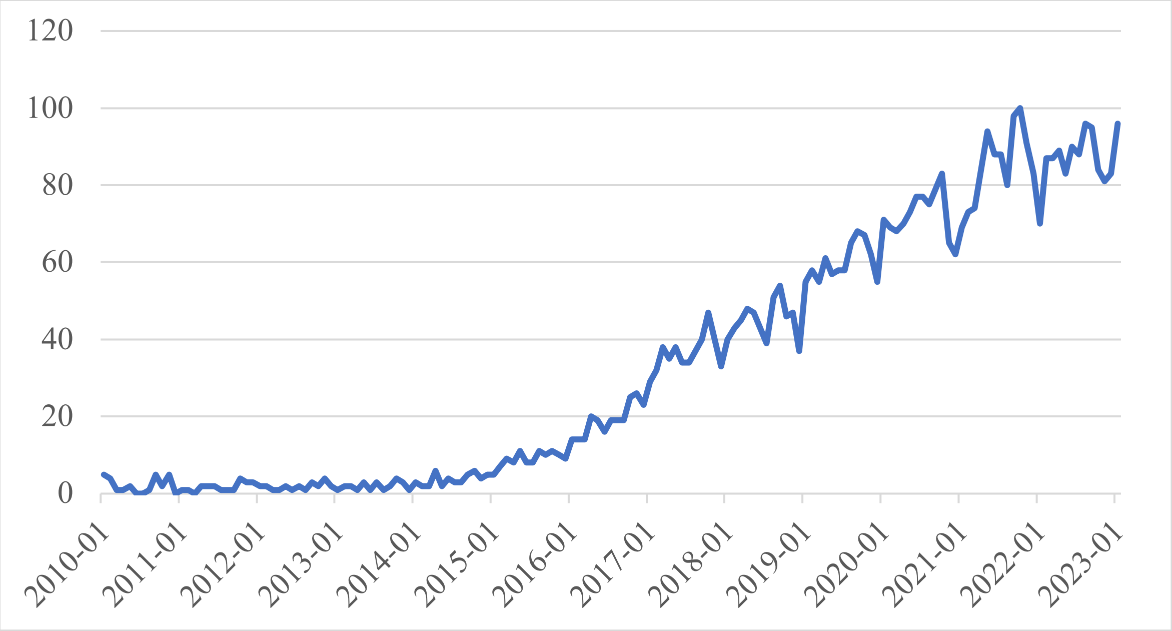
(b) Coarsened Exact Matching (CEM)



Appendix Figure B4: Corporations' Response to GDPR Enforcement by Intrafirm Data Interoperability. The plotted estimates are the treatment effects and 95% confidence intervals of GDPR enforcement over time. Time 0 indicates the year of GDPR enforcement (2018). Other times are the number of years relative to Time 0. Controls include corporation fixed effects and year fixed effects. Robust standard errors are clustered at the corporation level. The subfigures compare corporations responding "No" (top) and "Yes" (bottom) to the survey question: *"Do you use services interfaces or APIs to publish departmental (e.g., not application) information between departments and/or functional areas?"*



Appendix Figure B5: Google Trends Results for Digital Transformation in the United States. This figure shows Google search interest in "digital transformation" in the United States region between 2010 and 2022.



Appendix Table C1: List of Corporations and Their Responses to the Internal Data API Survey Question. This table lists the corporations by their response to the survey question on internal data APIs.

Yes (=1)	**No (=0)**
American Express Company	*American International Group, Inc.*
Bank of America Corporation	*Assurant, Inc.*
Capital One Financial Corporation	*Eaton Vance Corp.*
Centene Corporation	*Markel Corporation*
Citigroup Inc.	*MetLife, Inc.*
JPMorgan Chase & Co.	*The Hartford Financial Services Group, Inc.*
Mastercard Incorporated	*Wells Fargo & Company*
MoneyGram International, Inc.	
Prudential Financial, Inc.	
Chubb Limited	
Fifth Third Bancorp	
Lincoln National Corporation	
Morgan Stanley	
The Bank of New York Mellon Corporation	
The Charles Schwab Corporation	
The Goldman Sachs Group, Inc.	
The Travelers Companies, Inc.	
UnitedHealth Group Incorporated	

Appendix Table C2: Descriptive Statistics of Reported vs. Imputed Log Revenues in 2016. This table presents the descriptive statistics on establishment log revenues in 2016 for (1) reported non-missing values in the original CITDB data, (2) imputed values for establishments with non-missing revenues, and (3) imputed values for all establishments in the CITDB data. The correlation between (1) and (2) is 0.9685.

	Mean	SD	P5	P10	P50	P75	P90	P95	#Obs.
Original	17.595	2.664	13.816	13.816	15.607	17.312	19.275	21.154	866
Imputed (Orig)	17.600	2.613	13.800	14.468	15.541	17.321	19.204	21.222	866
Imputed (All)	14.922	1.149	13.773	13.775	13.814	14.871	15.200	16.212	36500

Appendix Table C3: Using 10-K Texts to Determine Whether Europe is the Primary Foreign Market. This table summarizes five corporations' 10-K texts in 2015, which enabled us to determine whether Europe is the primary market where each corporation operates outside the U.S.

(a) Significant Revenue Contribution from the European Market

<table>
<tr><td>MoneyGram International, Inc.</td><td>Keyword Counts: Germany/France/Italy/Spain (10) United Kingdom/Ireland (10) Europe (27)</td></tr>
<tr><td colspan="2">*"The Global Funds Transfer segment is our primary revenue driver, providing money transfer services and bill payment services primarily to unbanked and underbanked consumers… As of December 31, 2015, our money transfer agent network had over 350,000 locations, with growth of 3 percent compared to 2014. Our agent network includes agents such as international post offices, formal and alternative financial institutions as well as large and small retailers. Additionally,* ***we have Company-operated retail locations in the U.S. and Western Europe****."* (Management's Discussion and Analysis of Financial Condition and Results of Operations).
"MoneyGram offers products and services under its two reporting segments: Global Funds Transfer ("GFT") and Financial Paper Products ("FPP"). The GFT segment provides global money transfer services and bill payment services to consumers. We primarily offer services through third-party agents, including retail chains, independent retailers, post offices and other financial institutions. We also offer Digital/Self-Service solutions such as moneygram.com, mobile solutions, account deposit and kiosk-based services. Additionally, we have Company-operated retail locations in the U.S. and Western Europe… ***Our primary overseas operating subsidiary, MoneyGram International Ltd., is a licensed payment institution in the United Kingdom, enabling us to offer our money transfer service in the European Economic Area****."* (Notes to Consolidated Financial Statements)</td></tr>
<tr><td>Mastercard Incorporated</td><td>Keyword Counts: United Kingdom/Ireland (7) Europe (39)</td></tr>
<tr><td colspan="2">*"The European Commission issued a Statement of Objections in July 2015 related to* ***our interregional interchange fees and central acquiring rules within the European Economic Area****… In addition, due to the European Court of Justice's recent invalidation of the Safe Harbor treaty,* ***we may be subject to enhanced compliance and operational requirements in the European Union****."* (Risk Factors)
"In July 2015, the European Commission issued a Statement of Objections related to MasterCard's interregional interchange fees and central acquiring rules within the European Economic Area… In the United Kingdom, beginning in May 2012, a number of retailers filed claims against MasterCard seeking damages for alleged anti-competitive conduct with respect to MasterCard's cross-border interchange fees and its U.K. and Ireland domestic interchange fees. More than 30 different retailers have filed claims or threatened litigation. Approximately 30 additional merchants have filed or threatened litigation with respect to interchange rates in Europe ("Pan-European claimants"). Although the U.K. and Pan-European claimants have not quantified the full extent of their compensatory and punitive damages, their purported damages exceed $2 billion… ***During the fourth quarter of 2015, the Company designated its €1.65 billion euro-denominated debt as a net investment hedge for a portion of its net investment in European foreign operations.****"* (Notes to Consolidated Financial Statements)</td></tr>
<tr><td>Eaton Vance Corp.</td><td>Keyword Counts: Europe (11) United Kingdom/Ireland (15)</td></tr>
<tr><td colspan="2">*Eaton Vance Management (International) Limited ("EVMI"), a wholly owned financial services company registered under the Financial Services and Market Act in the United Kingdom,* ***markets our products and services in Europe and certain other international markets****… We are headquartered in Boston, Massachusetts and also maintain offices in Atlanta, Georgia; Minneapolis, Minnesota; New York, New York; Seattle, Washington; Westport, Connecticut; London, England; Singapore; and Sydney, Australia.* ***Our sales representatives operate throughout the United States and in the United Kingdom, Europe****, Asia Pacific and Latin America. We are represented in the Middle East through an agreement*</td></tr>
</table>

with a third-party distributor… The Eaton Vance International (Ireland) Funds Plc. are Undertakings for Collective Investments in Transferable Securities (“UCITS”) funds domiciled in Ireland and sold by EVMI through certain intermediaries, and in some cases directly, ***to investors who are citizens of the United Kingdom, member nations of the European Union*** *and other countries outside the United States. The Eaton Vance International (Cayman Islands) Funds are Cayman Island-domiciled funds sold by EVMI and EVD through intermediaries to non-U.S. investors.* (Business) ***Our operations in the United Kingdom, the European Economic Area***, *Australia and Singapore are subject to significant compliance, disclosure and other obligations. We incur additional costs to satisfy the requirements of the European Union Directive on Undertakings for Collective Investments in Transferable Securities and the Alternative Investment Fund Managers Directive (together, the “Directives”). The Directives may also limit our operating flexibility and impact our ability to expand in European markets.* (Risk Factors) *Eaton Vance Corp. and its subsidiaries (the “Company”) manage investment funds and provide investment management and advisory services to* ***high-net-worth individuals and institutions in the United States, Europe*** *and certain other international markets.* (Notes to Consolidated Financial Statements)

(b) Insignificant Revenue Contribution from the European Market

UnitedHealth Group Incorporated	Keyword Counts: United Kingdom/Ireland (0) Germany/France/Italy/Spain (0) Europe (0)
Prudential Financial, Inc.	Keyword Counts: **Japan (212)** United Kingdom/Ireland (14) Germany/France/Italy/Spain (11) Europe (16)
Prudential Financial, Inc., a financial services leader with approximately $1.184 trillion of assets under management as of December 31, 2015, ***has operations in the United States, Asia****, Europe and Latin America. Through our subsidiaries and affiliates, we offer a wide array of financial products and services, including life insurance, annuities, retirement-related services, mutual funds and investment management… For our Asset Management segment, which includes our international investment operations,* ***as of December 31, 2015, we lease two home offices located in Japan and Taiwan****.* (Business) *Our International Insurance segment manufactures and distributes individual life insurance, retirement and related products, including certain health products with fixed benefits.* ***We provide these products to the broad middle income and mass affluent markets across Japan*** *through multiple distribution channels including banks, independent agencies and Life Consultants associated with our Gibraltar Life Insurance Company, Ltd. (“Gibraltar Life”) operations. We also provide similar products to the mass affluent and affluent markets through* ***our Life Planner operations in Japan, Korea and other countries outside the U.S., including Taiwan, Italy, Brazil, Argentina, Poland and Mexico****… For the year ended December 31, 2015, our* ***Life Planner and Gibraltar Life operations in Japan represented 37% and 51%, respectively, of the net premiums, policy charges and fee income of the International Insurance segment*** *and, in aggregate, represented* ***36% of the net premiums****, policy charges and fee income of Prudential Financial, translated on the basis of weighted average monthly exchange rates…* (International Insurance Division)	

Appendix Table C4: Descriptive Statistics of Aggregated Data. This table shows descriptive statistics on the aggregate data after averaging the establishment-level variables to the corporation level.

Variable	Mean	SD	P5	P25	P50	P75	P95	# Obs.
Ln(Revenue)	16.121	1.132	14.597	15.142	15.948	16.977	17.888	120
TREAT	0.583	0.495	0.000	0.000	1.000	1.000	1.000	120
API	0.708	0.456	0.000	0.000	1.000	1.000	1.000	120
STD	0.029	0.011	0.003	0.024	0.027	0.036	0.047	90
Log #PCs	3.567	0.548	2.900	3.083	3.455	3.972	4.537	120
Cloud Presence	0.763	0.256	0.303	0.565	0.886	0.969	0.995	120
ΔRevenue (2016-17)	0.035	0.066	-0.092	-0.002	0.041	0.072	0.144	120

Appendix Table C5: Effects of GDPR Enforcement on Establishment Revenue by Subsample Splits. The tables show regression results on establishment-year panel data, in subsamples split by internal data APIs in panel (a) and by median standardization of identity and access management software in panel (b). Confidence intervals are shown in round brackets (instead of standard errors), estimated using wild cluster bootstrap with N=9999 which generates more conservative CI estimates than cluster robust standard errors when the data structure consists of a small number of large clusters. P-values are shown in square brackets. *∗∗∗p < 0.01; ∗∗p < 0.05; ∗p < 0.1.*

(a) Sample Split by Internal Data APIs

Dependent Variable	Ln(Revenue)	
Sample: $I(API = 1)$	(1) No	(2) Yes
TREAT × POST	-0.115***	-0.030**
	(-0.125, -0.094)	(-0.093, -0.007)
	[0.001]	[0.032]
Year FE	Y	Y
Estab. FE	Y	Y
Obs.	27,630	58,925
# Establishments	5,526	11,785
# Years	5	5

(b) Sample Split by Standardization of IAM Software

Dependent Variable	Ln(Revenue)	
Sample: $I(STD \geq P50)$	(1) No	(2) Yes
TREAT × POST	-0.085**	-0.005
	(-0.158, -0.004)	(-0.216, 0.371)
	[0.039]	[0.120]
Year FE	Y	Y
Estab. FE	Y	Y
Obs.	27,430	38,310
# Establishments	5,486	7,662
# Years	5	5

Appendix Table C6: Corporation-Level Analyses – Difference-in-Differences and Synthetic DID Estimates. The tables show regression results on corporation-year panel data, in subsamples split by internal data APIs (measured at the corporation level). Robust standard errors are clustered at the corporation level for the DID estimates. $***p < 0.01; **p < 0.05; *p < 0.1$.

Dependent Variable	Log (Revenue)			
	(1)	(2)	(3)	(4)
$I(API = 1)$	No DID	No SynthDID	Yes DID	Yes SynthDID
TREAT × POST	-0.094*** (0.021)	-0.094*** (0.017)	-0.024 (0.045)	-0.007 (0.055)
Year FE + Corporation FE	Y	Y	Y	Y
Observations	35	35	85	85
# Corporations	7	7	17	17
# Years	5	5	5	5