Data Capital and Innovation: Evidence from Connected Devices*

Tarikua Erda† Ryan Gilland‡ Ben Lahey§

July 6, 2026

Click here for the latest version

Abstract

Consumer data is a valuable form of intangible capital, but goes unmeasured in standard firm-level datasets. Applying natural language processing to the user manuals of the universe of wireless devices sold in the U.S., we construct the first measure of firms' data capital: each firm's portfolio of connected "smart" devices that collect and transmit user data through apps, accounts, and cloud servers. Data-capable public firms are large, highly R&D-intensive, and have dominated patenting in machine learning and AI, domains where data is a core input. Using a triple-difference design around the European General Data Protection Regulation's restrictions on personal data, we find substantial declines in R&D investment and patenting among consumer-facing data-capable firms, with innovation slowdowns primarily concentrated in data-intensive AI domains. Financial markets durably marked down the valuations for GDPR-exposed firms with larger smart consumer gadget portfolios, reflecting the decline in expected future rents from personal data. Our approach offers a new, scalable measure of data capital with direct implications for firm performance and innovation in the AI era.

Keywords: Data Economy; Intangible Capital; Innovation; R&D; Internet of Things; AI; Corporate Finance
JEL Codes: E22; G30; O34; L86; K20

*We thank Anton Babrov, Olalekan Bello, Michael Ewens, Xavier Giroud, Deepak Hegde, Thomas Philippon, Rob Seamans, and James Traina for helpful discussions. Support from the Alfred P Sloan Foundation is gratefully acknowledged. The authors have no financial or other ties related to the research project. All errors are ours.

†Erda: New York University, tarikua.erda@nyu.edu.
‡Gilland: New York University, rg4795@nyu.edu.
§Lahey: New York University, bpl9631@nyu.edu.

1 Introduction

The ability to collect, transmit, and analyze user data has become a central form of investment for modern firms. User data shapes product improvement, differentiation, and pricing, and underpins the development of AI and prediction technologies, with venture capital firms also pointing to the quality and uniqueness of startups' proprietary data as their biggest competitive edge (Szkutak, 2025). Data has been recognized in the academic literature as a core intangible asset and modeled as a driver of firm dynamics (Farboodi et al., 2019), merger activity (de Cornière & Taylor, 2024), the macroeconomy (Farboodi & Veldkamp, 2025), and allocative efficiency under different property rights regimes (Jones & Tonetti, 2020). Yet, we have relatively little empirical evidence on how exactly data factors into firm performance and innovation, which is the subject of this paper.

Quantifying the value of data capability as an intangible asset has been difficult, primarily due to measurement challenges: standard proxies for intangible investment (accounting intangibles, patents, job postings, or web-based measures) do not directly observe firms' ability to collect data. Existing "gold standard" firm surveys from the US Census Bureau can tell us whether firms report *using* data, cloud computing, or analytics, but they remain coarse, self-reported, and infrequent.[1]

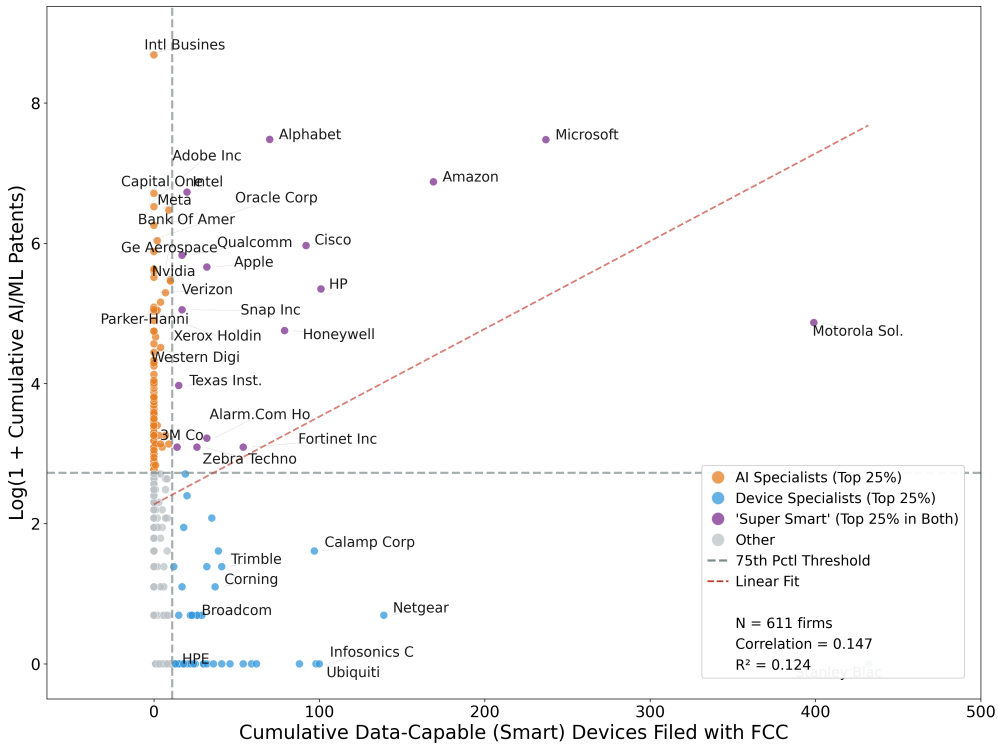
In this paper, we use an untapped administrative dataset on wireless products and connected devices ("smart" products that connect to apps or the cloud and continuously generate, process, and transmit user data) to develop a novel measure of firms' data capability. Using this measure, we document new facts about the role of data in firm performance and innovation. We also exploit a data collection cost shock created by the European Union's General Data Protection Regulation ("GDPR"), which restricted key forms of personal data transmission and processing for firms with customers in the European Economic Area. Data-capable firms selling consumer-facing smart devices cut R&D investment and experience sharp declines in AI and machine learning patenting relative to those selling industrial smart products. Tobin's Q falls for consumer gadget producers with larger smart device portfolios as well, reflecting the declining expected rents from data for these firms. Overall, our findings show that data is an accumulated, depreciable intangible capital and a productive input into the innovation pipeline.

Connected devices make a useful setting to study data capital for two reasons. First, the booming growth in these devices defines the modern, Internet of Things ("IoT") era of

[1]For example, the Annual Survey of Manufactures (ASM) and the Management and Organizational Practices Survey (MOPS) include valuable questions on IT and management practices, but they do not measure product-level data capability and are only available for a handful of sampled firms and years.

consumer electronics. The enterprise IoT market stood at $269 billion in 2023, with recent estimates citing more than 20 billion connected devices globally and more than 120 zettabytes of global data creation/replication that is projected to grow further in the rest of this decade.[2] Second, the economics of device-generated data differ from the web-tracking setting that dominates much of the existing privacy literature.[3] Such measures of data intensity built on third-party tracking are intertwined with browsers, consent banners, and platform rules. In contrast, as part of ordinary product use, connected devices generate first-party data: data that is proprietary, tied to an installed base, and able to be reused to improve products, refine predictive analytics, and develop complementary products. Measuring the value of data capability embodied in connected devices is therefore a natural and timely step toward understanding data as intangible capital in the AI era. Figure 1 previews this link using our data on smart devices and patent records: there is a positive association between the number of data-capable devices a firm offers and its patenting in AI and machine learning, with a few 'super smart' firms highly active in both domains clustered in the upper right.

Figure 1: Smart Devices and Patenting in AI/Machine Learning Domains



For intuition, consider a robot vacuum. The vacuum generates a steady stream of usage and performance data (cleaning patterns, navigation maps, error logs, battery wear),

[2]See: BusinessWire.com (2021); IOT-Analytics.com (2025)

[3]For instance, see Aridor et al. (2023); Liu (2024).

often routed through a companion app and cloud service. The firm can use these data to improve navigation and reliability, inform product innovation, and support add-on services such as software features and replacement parts. The installed, tangible hardware generating this intangible data stream thus enables value creation beyond the initial sale.

We observe connected devices sold in US markets using the mandatory device authorization filings from the Federal Communications Commission ("FCC"), which are required for virtually any wireless device sold in the United States. These filings provide a census of all radio frequency-emitting devices, along with rich product documentation including, notably, user manuals that describe device connectivity, companion applications, and cloud interactions. They are uniquely useful for our purposes because they provide revealed, engineering-focused, product-level capabilities at the point of commercialization. They are thus free from selection concerns that may arise from voluntary self-reporting contexts where firm strategic behavior may matter, and they offer a more reliable paper trail to track product innovation than patents, whose contents may not make it to market.

To identify devices with data-collection capability, we apply natural language processing ("NLP") methods to user manual texts, focusing on features such as companion-app connectivity and cloud data transmission. Aggregating across devices, we construct a firm-year measure of data capability that captures whether a firm owns an FCC certification for a data-capable device (and how many).[4] We primarily focus on devices filed between 2007, when smartphones and apps were introduced, and 2019, the last year for which the FCC website provides detailed and comprehensive coverage.

We document that data capability is associated with more prolific new product filings in subsequent years. FCC grantees that have previously commercialized data-capable devices are more likely to file new products in a given year and introduce greater numbers of new devices when they do, consistent with data-enabled products anchoring follow-on innovation.

We then use Compustat to characterize the public firms selling wireless devices, finding that data capability has increasingly been concentrated among larger, R&D-intensive and innovative firms over the course of the IoT boom. Public firms that sell data-capable products are generally large in assets and higher valued. They also spend much more on R&D relative to non-data-capable public firms. They have come to dominate patenting activity in AI and machine learning technologies, where data is a critical input. These patterns, which also hold after sector adjustment, indicate that data capability is closely tied to innovative output, R&D strategy, and market position within industries, not merely across them.[5]

[4]We validate this measure by showing that firms we classify as data-capable are substantially more likely to file deal narrative texts with the SEC that mention collection and usage of data.

[5]Among early-stage private firms, bringing a data-capable device to market was associated with a higher

Finally, we use a triple-difference design to study the impact of the GDPR's privacy regulation shock that increased firms' compliance costs for using first-party data generated from devices. In particular, we examine how the outcomes of data-collecting firms with exposure to European customers change after GDPR is adopted in 2016 relative to non-data-capable firms with EU exposure. This approach helps to isolate the GDPR's impact from confounders like broader differences in firm technology, managerial quality, and growth opportunities.

We document that EU-exposed data-capable consumer firms reduced R&D activity and lost value after GDPR. Consumer-gadget makers cut R&D spending by roughly 40% relative to control firms, and their new patent applications in AI/ML areas fell by 31%. Under an independent, text-based classification of AI patents, the contraction concentrates in computer vision and speech recognition, the applications most dependent on the personal sensory data that consumer devices collect. The financial responses also scale with the size of the data-capital stock: each additional data-capable device in a firm's pre-GDPR portfolio is associated with a further 3% decline in R&D and a 2.3% decline in Tobin's Q. In contrast, firms making smart devices for industrial use, whose data collection falls largely outside GDPR's personal-data scope, show no such declines and if anything increase their AI/ML patenting, consistent with GDPR selectively disrupting innovation that depends on personal consumer data. Together with the descriptive evidence, these results suggest that first-party data is a productive input into innovation: when regulation restricts its collection and reuse, investment and patenting contract in the domains most reliant on data. In ongoing work, we are refining and expanding the grantee-to-firm matching to gain a larger and richer sample.

This paper makes several contributions. First, we provide novel evidence on the value of data capital for firm outcomes in the context of connected devices, which constitute a substantial share of data generation in the current digital age. Previous work had documented sales declines of high data-intensity firms after GDPR (see Quan (2024); Liu (2024)) in the context of website data. We extend this to first-party data collection via connected devices. Our finding that AI/ML patenting declines sharply among EU-exposed data-capable consumer firms adds to a small but growing body of evidence that privacy regulation affects the direction and pace of innovation (see Blind et al. (2024); Giorgio et al. (2024)). Our contribution is to identify this effect at the firm level using product-based measures of data capability, and to show that the decline is concentrated among consumer-facing firms whose data collection falls squarely within GDPR's personal-data scope. The consumer/non-consumer contrast, in which industrial device makers are unaffected or even benefit, strengthens the

likelihood of a successful exit, particularly through IPO, though this advantage faded by the mid-2010s as IoT technologies diffused (Appendix A.4).

case that the mechanism runs through personal data specifically, not through broader regulatory burden. In this way, our paper also provides empirical evidence consistent with data being a non-rival, growth-enhancing asset as modeled in previous theoretical work (see Jones and Tonetti (2020); Farboodi and Veldkamp (2025)).

Second, our work contributes to a growing literature on intangible capital (Crouzet & Eberly, 2023; Crouzet et al., 2022) by providing a direct, product-level measure of one of its least-observed components: data. Existing empirical proxies for firm data intensity typically rely on disclosures like 10-K filings (Boroomand et al., 2022), online tracking measures (Goldberg et al., 2024; Aridor et al., 2023), IT and data-processing activity from cloud usage (Demirer et al., 2024), or hiring and data-related patents (Quan, 2024). While informative, this prior work leaves unobserved firms' first-party data generation outside of websites. Our use of FCC authorization filings allows us to contribute a new and comprehensive measure of data capability that is embodied in physical hardware products. Because FCC authorization occurs at or near commercialization and is produced for engineering compliance rather than investor communication, the resulting measure captures deployable data-collection infrastructure with a measurement error that is plausibly orthogonal to firms' strategic disclosure incentives.

In this respect, our methodology also adds to the literature using NLP to convert unstructured text into structured economic statistics to understand the economy, especially critical areas like artificial intelligence (Babina, Fedyk, He, & Hodson, 2024) and fundamental growth-drivers like innovation (Bergeaud, Jaffe, & Papanikolaou, 2025). Beyond allowing a new measure of data capability in connected devices, our use of user manual text opens new avenues for the study of product innovation and technology in the AI era.

2 Data and Measurement

2.1 Data

FCC Equipment Authorization Filings

Any wireless device marketed or imported into the United States must obtain authorization from the Federal Communications Commission (FCC) before entering commerce. This requirement applies broadly to equipment that emits radio-frequency (RF) energy, including consumer electronics, industrial sensors, medical devices, and Internet-of-Things (IoT) products. Device authorization requires submission of an application (Form 731) identifying the

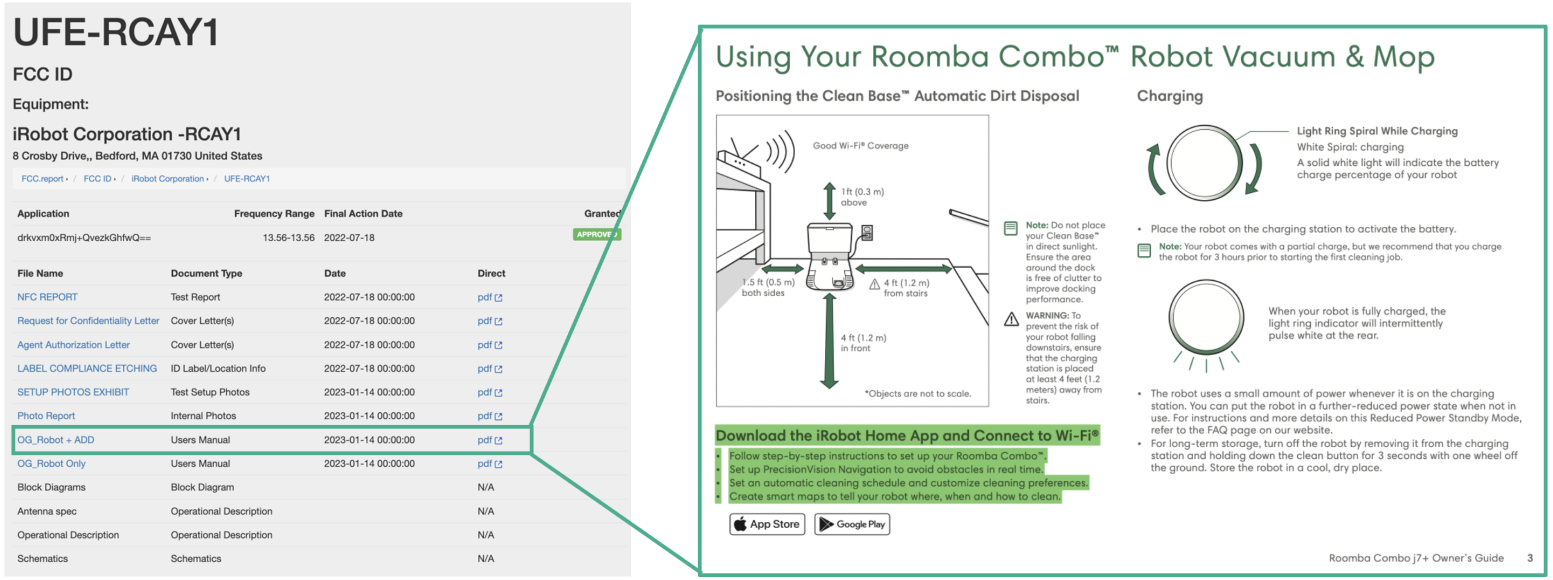
legally responsible party (the "grantee"), the device's technical characteristics, applicable rule parts, and the authorization date. The grantee need not be the physical manufacturer; it may be a brand owner, importer, or distributor, provided it assumes legal responsibility for compliance. For foreign manufacturers, FCC rules require a U.S.-based agent for service of process, ensuring each authorized device is linked to a U.S.-jurisdictional entity. Because authorization precedes market access, these filings are mandatory and independent of firms' financial reporting considerations, performance, or strategic disclosure choices.

We build a device-level FCC dataset by scraping the metadata and application documentation on user manuals associated with device filings.[6] These filings allow us to observe the technical exhibits submitted during the approval process. They also include information on rule parts, frequency bands, and transmission technologies. These materials allow us to infer the economically meaningful device features that distinguish simple RF devices (e.g., garage door openers) from data-capable connected products (e.g., smart wearables and home appliances), such as whether a product supports persistent connectivity, data transmission, or companion mobile applications.

As an illustrative example, Figure 2 shows the FCC equipment authorization records for the Roomba, an autonomous robotic vacuum device, from fcc.report. The top left panel shows an iRobot filing in the FCC Equipment Authorization System, including the exhibit list that typically contains test reports and publicly posted user manuals. The top right panel zooms in on a page from the Roomba's user guide that instructs users to download a companion smartphone app and connect the device to Wi-Fi. We use NLP to detect app-companion and connectivity language, systematically identifying connected devices that likely transmit user or sensor data.

[6] These were accessed from https://fccid.io/ and https://fcc.report/.

Figure 2: FCC record example for the Roomba (FCC ID UFE-RCAY1)



UFE-RCAY1

FCC ID

Equipment:

iRobot Corporation -RCAY1

8 Crosby Drive,, Bedford, MA 01730 United States

FCC.report › / FCC ID › / iRobot Corporation › / UFE-RCAY1

Application	Frequency Range	Final Action Date	Grantee
drkvxm0xRmj+QvezkGhfwQ==	13.56-13.56	2022-07-18	APPROVE

File Name	Document Type	Date	Direct
NFC REPORT	Test Report	2022-07-18 00:00:00	pdf
Request for Confidentiality Letter	Cover Letter(s)	2022-07-18 00:00:00	pdf
Agent Authorization Letter	Cover Letter(s)	2022-07-18 00:00:00	pdf
LABEL COMPLIANCE ETCHING	ID Label/Location Info	2022-07-18 00:00:00	pdf
SETUP PHOTOS EXHIBIT	Test Setup Photos	2023-01-14 00:00:00	pdf
Photo Report	Internal Photos	2023-01-14 00:00:00	pdf
OG_Robot + ADD	Users Manual	2023-01-14 00:00:00	pdf
OG_Robot Only	Users Manual	2023-01-14 00:00:00	pdf
Block Diagrams	Block Diagram		N/A
Antenna spec	Operational Description		N/A
Operational Description	Operational Description		N/A
Schematics	Schematics		N/A

Using Your Roomba Combo™ Robot Vacuum & Mop

Positioning the Clean Base™ Automatic Dirt Disposal

Good Wi-Fi® Coverage

1 ft (0.3 m) above

1.5 ft (0.5 m) both sides

4 ft (1.2 m) from stairs

4 ft (1.2 m) in front

*Objects are not to scale.

Note: Do not place your Clean Base™ in direct sunlight. Ensure the area around the dock is free of clutter to improve docking performance.

WARNING: To prevent the risk of your robot falling downstairs, ensure that the charging station is placed at least 4 feet (1.2 meters) away from stairs.

Download the iRobot Home App and Connect to Wi-Fi®

- Follow step-by-step instructions to set up your Roomba Combo™.
- Set up PrecisionVision Navigation to avoid obstacles in real time.
- Set an automatic cleaning schedule and customize cleaning preferences.
- Create smart maps to tell your robot where, when and how to clean.

App Store Google Play

Charging

Light Ring Spiral While Charging

White Spiral: charging

A solid white light will indicate the battery charge percentage of your robot

- Place the robot on the charging station to activate the battery.

Note: Your robot comes with a partial charge, but we recommend that you charge the robot for 3 hours prior to starting the first cleaning job.

When your robot is fully charged, the light ring indicator will intermittently pulse white at the rear.

- The robot uses a small amount of power whenever it is on the charging station. You can put the robot in a further-reduced power state when not in use. For instructions and more details on this Reduced Power Standby Mode, refer to the FAQ page on our website.
- For long-term storage, turn off the robot by removing it from the charging station and holding down the clean button for 3 seconds with one wheel off the ground. Store the robot in a cool, dry place.

Roomba Combo j7+ Owner's Guide 3

Application Details:

Equipment	Automated robot vacuum
FRN	0015256720
Grantee Code	UFE
Product Code	-RCAY1
Applicant Business	iRobot Corporation
Business Address	8 Crosby Drive, , Bedford,Massachusetts United States 01730
TCB Scope	A1: Low Power Transmitters below 1 GHz (except Spread Spectrum), Unintentional Radiators, EAS (Part 11) & Consumer ISM devices
TCB Email	LST.FRE.CertNotice@ul.com
Applicant	Stephen Pallotta – Principal Compliance Engineer
Applicant Phone	
Applicant Fax	781-430-3001
Applicant Email	spallotta@irobot.com
Applicant MailStop	M/S 12-1
Test Firm	UL LLC
Test Firm Contact	Steven Thompson
Test Firm Phone	
Test Firm Fax	
Test Firm Email	steven.d.thompson@ul.com
Certified By	Stephen Pallotta –

Firm Data

For our analysis of the effects of connected devices on firm cash flows, we obtain regulatory financial filings from Compustat. As a majority of FCC filings are submitted by firms located outside of the United States, we target quarterly financial data for the entirety of firms in the Compustat universe, harmonizing quantities reported in historical local currencies to 2024 USD. We focus on a handful of core measures broadly related to firm performance and innovative output: revenue, R&D expenditures, R&D intensity (defined as R&D normalized by revenue), and market capitalization. We additionally compute Tobin's Q for each firm, using the simplified computation discussed in (Chung & Pruitt, 1994). Given we are analyzing a sample of firms that must first be fuzzy-matched across multiple data sources, we prefer to build this measure in a manner that best preserves our linked sample.

We incorporate firm customer segment reporting to build a measure of exposure to markets subject to European Union regulation. SEC reporting rules require firms to disclose revenues from individual customers that contribute more than 10 percent of consolidated revenue (Board, 2026). Disclosure requirements across other broad categories are more opaque. For instance, the Financial Accounting Standards Board's (FASB) Accounting Standards Codification (ASC) 280-10-55-51 states that public entities must report "revenues from external customers attributed to the public entity's country of domicile and attributed to all foreign countries in total from which the public entity derives revenues. If revenues from external customers attributed to an individual foreign country are material, those revenues shall be disclosed separately." Here "material" is not explicitly defined and thus geographic disclosures are somewhat at the discretion of management. Major accounting firms such as Deloitte and PricewaterhouseCoopers take the perspective that, in line with individual customers, firms should report countries which exceed a contribution of 10 percent to consolidated revenues or long-lived assets[7].

Patent Filings

We augment our Compustat sample with patent grant data from the United States Patent and Trademark Office (USPTO). We use the December 7, 2025 revision of the Center for Research in Security Prices (CRSP) to USPTO patent grants crosswalk introduced in Kogan, Papanikolaou, Seru, and Stoffman (2017) as our baseline sample of patents. In this sample, we identify patents as being AI/ML-adjacent using granular USPTO Cooperative Patent Classification (CPC) codes assigned to patents.

[7]The perspectives of Deloitte and PricewaterhouseCoopers can be viewed through these links.

In our analyses we emphasize the date a patent was first filed with the USPTO, rather than the date it was granted, to account for the lag in the time between filing and approval. Each patent's citation weight is computed as the ratio of its forward citations to the average forward citations received by all patents in the same CPC3 technology class and grant year, following the approach in Hall et al. (2005). We aggregate these weights to the firm-year level by summing across all patents filed by a given firm in a given year. This weighting adjusts for well-documented differences in citation propensities across technology fields and over time, ensuring that a patent in a sparsely cited class is not mechanically undervalued relative to one in a heavily cited class.

Narratives from SEC Acquisition Filings

For the purpose of validating our FCC-based measure of data capability, we use SEC filings, typically 10-K annual reports or 8-K current reports, which are a required element of acquisitions completed by publicly traded firms. These filings contain both structured purchase price allocations that value identifiable intangibles (e.g., technology, customer relationships, goodwill) and narrative text describing the deal's strategic rationale. We use the 627 acquisitions (2007–2017) compiled by Ewens et al. (2025), which extracts intangible-asset valuations from these filings. The narrative sections often characterize targets' assets in ways not captured by line items, including references to data capabilities, user engagement, monetization potential, and technology infrastructure.

2.2 Developing and Validating our Measure of Data Capability

Approach for Identifying Data-Capable Devices: Our primary measure comes from a large language model (Anthropic's Claude Opus 4.8) that reads the full manual and records a structured profile of the device through a constrained tool call. For every device the model returns a fixed set of indicators: whether the device is operated through a mobile companion app, whether it requires a user account, whether it connects to a cloud service, whether it ships with desktop configuration software, and whether the manual describes an explicit pairing or setup flow. For each indicator it sets to one, the model must also return a verbatim quotation from the manual supporting that classification, so any positive flag in the dataset can be checked against the underlying text. The prompt instructs the model to judge only from the manual, to leave an indicator at zero when the text is silent rather than infer from brand knowledge, and to separate mobile companion apps from desktop configuration utilities. A device is data-capable under our narrow measure if its manual documents a mobile

app, a user account, or a cloud service. These are the indicators that map most directly into the collection of personal data regulated under GDPR, and the narrow measure is the one we use throughout the paper. A broad measure additionally counts desktop software and pairing functionality. Appendix A.1 describes the extraction schema, the prompt rules, the processing pipeline, and the checks we ran on the model's output.

As a complementary measure, we also use an algorithm that searches for textual patterns indicating companion-app connectivity and cloud data transmission. Devices are classified as "data capable" if their user manuals contain either (i) any explicit reference to companion apps, app stores, or brand-specific app names (e.g., "mobile app," "download from Google Play," "Ring app"), or (ii) at least two references to secondary indicators such as remote access functionality, push notifications, account creation, cloud storage, or QR-code-based setup. We additionally flag devices mentioning data synchronization, usage tracking, or server-based storage. Documents that appear to be test reports rather than user manuals are filtered out prior to classification. A more detailed summary on this approach is provided in Appendix section A.2.

Appendix Figure A.1 displays the number of approved FCC filings by grantee country and state for filings concerning data-collecting devices over the 2007-2025 period. Of the 60,077 filings we are able to map to grantee country, 20,534 (34.2 percent) were granted to firms headquartered in the United States. The remaining top five most prevalent jurisdictions with data-collecting device approvals were China and Hong Kong (31.0 and 4.8 percent), Taiwan (7.4 percent), South Korea (4.0 percent), and Japan (2.8 percent). Just outside is Germany (2.1 percent), an EU member. Decomposing the filings granted to firms headquartered in the United States, the most prevalent states are California (28.6 percent of US filings), New Jersey (18.7 percent), Florida (7.2 percent), Illinois (6.1 percent), and New York (4.1 percent). Thus device filings are highly concentrated among a handful of countries, primarily the US and China, and within the US, California.

Validation: We perform an internal check of the LLM classification approach using a second, independently trained, and more capable language model, Claude Fable 5. An independent audit of 500 manuals by Fable, oversampling devices on which the Opus 4.8 and the text-based classifications disagree, confirms 95 percent of the classification's positive calls; its residual errors are conservative, under-counting rather than over-counting data capability. The extraction pipeline is deterministic and reproduces the classification file exactly on re-runs (Appendix A.1).

We also use a separate, external resource: the texts of purchase price allocations in Securities Exchange Commission (SEC) filings. In particular, we examine whether acquisi-

tions involving companies with data-collecting devices exhibit different linguistic patterns in their SEC filings compared to acquisitions without such devices.

We proceed in three steps. First, we start with the Ewens et al. (2025) Purchase Price Allocation (PPA) dataset. We find a successful match to our FCC grantee database for 175 deals, where either the target or acquiring firm sold FCC-licensed devices. Of these, 31 deals were specifically linked to target firms.[8] Next, for each acquisition, we extract text excerpts from the corresponding SEC filings that mention "data" or related terms in the context of business combinations, purchase price allocations, and acquisition rationale. We then use NLP to identify filings mentioning keywords related to data strategy and use, including in the areas of user engagement and retention metrics, IoT and connected devices (e.g., wearables like Fitbits). Lastly, we compare the two groups of deals using Fisher's exact test for a binary indicator of *any* keyword presence.

Table 1: FCC grantees classified as likely data-collectors are substantially more likely to be associated with data use/data strategy related keywords in SEC acquisition filings

	Has Keyword	Total	%
Data-device deals	47	129	36.4%
No-data-device deals	9	46	19.6%
Difference			+16.9pp
Odds Ratio		2.36	
p-value (Fisher's exact)		0.0428**	

Notes: All deals (N=175) in which either the target or the acquirer is matched to an FCC grantee. Binary indicator equals 1 if any data-strategy keyword appears in SEC filing excerpts. Sample restricted to high-quality FCC-SEC matches (score $\geq$ 0.9). Deals whose matched grantees have no readable device manuals are classified as non-data.
*** p<0.01, ** p<0.05, * p<0.10

Our findings in Table 1 reveal that acquisitions involving data-device companies exhibit markedly higher mentions of data-related strategy keywords. Across the 175 deals in which either the target or the acquirer is matched to an FCC grantee, 36.4% of deals involving a data-device company have at least one data-strategy keyword in the deal narrative, compared with 19.6% of deals without one, an odds ratio of 2.4. The detailed breakdown of keyword types in Appendix Table A.1 shows the contrast is driven primarily by user engagement and retention language. These findings suggest that FCC data-device classifications capture meaningful variation in how acquirers discuss data assets and strategy in their SEC

[8]It is not surprising that we obtained a relatively much smaller set of matches to target firms: the SEC data primarily capture larger public firms while acquisition targets are likely to be smaller, private companies.

disclosures, and thus lend credence that our measurement of the data capability of firms is reasonable.

2.3 Sample Construction

Our data on FCC device filings contains a unique identifier associated with each grantee, but this identifier does not map to any relevant firm data. We thus use a matching algorithm to map firms recorded in FCC device filings to those in the CapitalIQ/Compustat database. Details of the matching algorithm are described in the appendix. Overall we match 4,861 (more than 16 percent) of the over 29,000 uniquely identified firms in the FCC device filings to those in CapitalIQ/Compustat.

We then use the latest CRSP to Compustat firm crosswalk supplied by CRSP to map patents to Compustat firms over time. We identify the patent assignee firm in our FCC-Compustat link for 1,195,227 (35 percent) of the 3,419,552 patents in our initial sample, of which 11,489 are classified as related to Artificial Intelligence ("AI") or machine learning ("ML"). We then build measures of patenting activity both broadly and within AI/ML, as well as weighted measures using the forward citations provided by (Kogan et al., 2017).

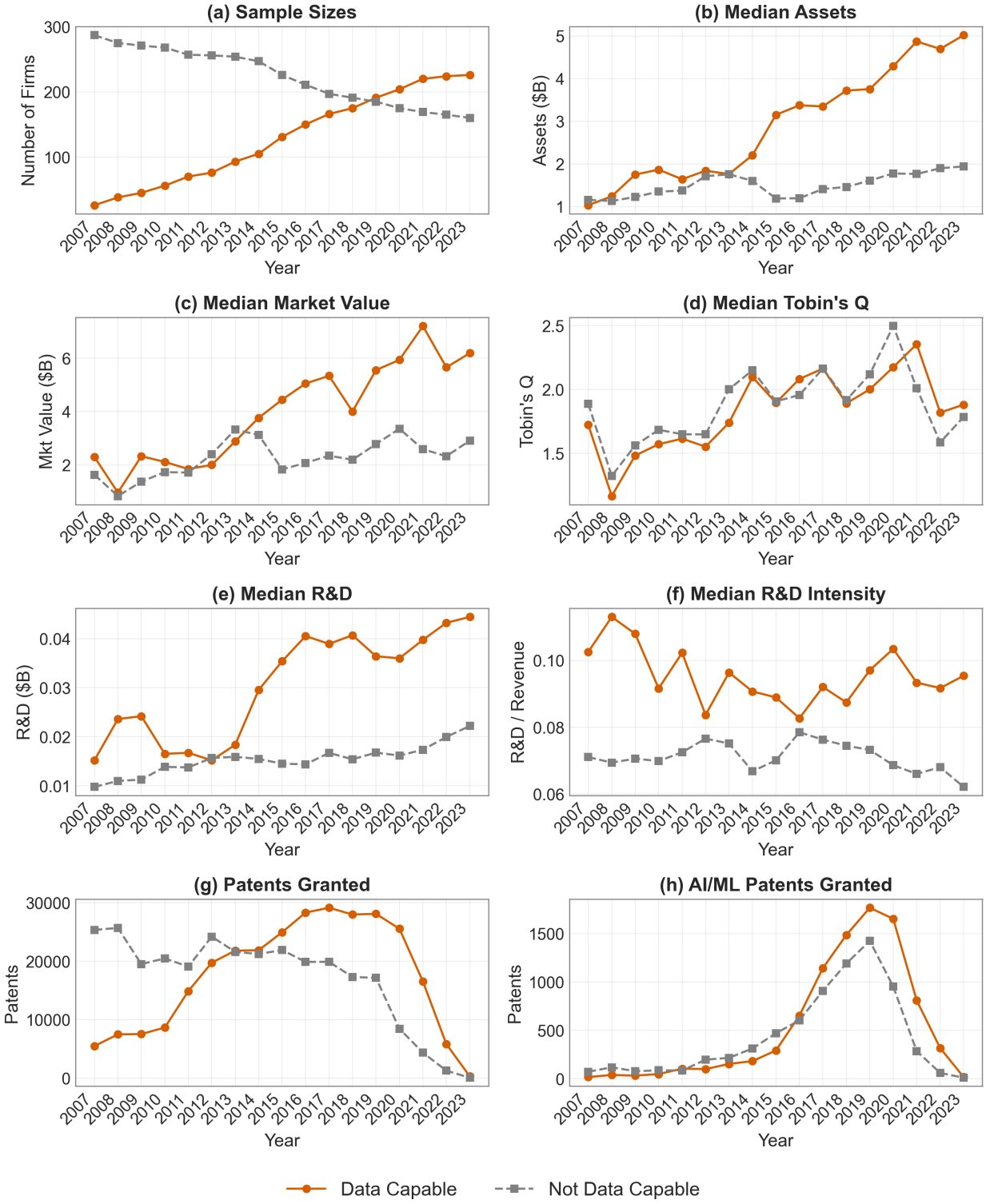
Lastly, we use reported customer segments to identify firms with exposure to the European Union. While there is expected to be some reporting bias in firm geographic segments given the lack of a hard reporting requirement, the loose "material" requirement implies the firms that do report splits do so because these geographies represent material concerns. We therefore define firms as being exposed to European Union markets (and thus European Union regulations) based on whether that firm reports revenue or asset exposures to a region or country situated in the European Union. For the purposes of our later analyses, we use sales reported as of the first quarter of 2016 (pre-GDPR adoption) to generate our indicator. Ultimately we identify roughly half (49.8 percent) of the FCC-linked device firms in our Compustat panel as being potentially exposed to GDPR.

3 Characterizing Data-Capable Public Firms

Figure 3 shows various summary statistics comparing public firms with at least one data-capable device against 'non-capable' firms with none.

Panel (a) shows that the number of data-capable public firms has grown steadily, overtaking non-capable firms by 2019 and reaching 226 versus 160 firms by 2023. This is consistent with a broad shift toward connected products since the proliferation of smartphones

Figure 3: Comparing data-capable versus non-capable public firms (2007-2023)



and app-enabled devices. Data-capable firms are substantially larger and more highly valued: their median assets and market value exceed those of non-capable firms, with the gap widening markedly from the early 2010s onward (Panels b and c). By 2023, median assets for data-capable firms are roughly $5 billion, compared with $1.9 billion for non-capable firms; median market values diverge similarly ($6.2 billion vs. $2.9 billion). MedEian Tobin's Q tracks relatively closely across the two groups for much of the sample period, though data-capable firms carry a modest premium in the later years of the panel (Panel d). These patterns generally hold in sector-adjusted comparisons as well (see Appendix Figure A.2).

Notably, data-capable firms also appear more innovative. Panels (e) and (f) show that data-capable firms spend more on R&D both in absolute terms and as a share of revenue, consistent with data from connected devices serving as an input into product improvement and development. The gap in R&D spending widens over time, while R&D intensity remains persistently higher for data-capable firms throughout the sample. Strikingly, data-capable firms have come to dominate patenting activity. Panel (g) shows that total patents granted to data-capable firms surpass those of non-capable firms by 2013 and pull away sharply, accounting for over 60% of all patenting by the years around GDPR. In AI and machine-learning patents (Panel h), data-capable firms overtake non-capable firms from 2016 onward and account for the majority of activity in the late 2010s. This is consistent with data capability and AI innovation being complements: firms that generate first-party data through their connected devices are disproportionately the ones investing in the prediction technologies that use it.

As shown in Figure 1 earlier in the paper, "smart device" innovation is highly concentrated in a small set of large, technologically sophisticated firms: companies that introduce many data-capable FCC devices also tend to be the ones with the deepest AI/ML patent portfolios. In general, we document a positive correlation of 0.15 between firms' data-capable device filings and their AI/ML patent output, suggesting that hardware-enabled data collection and artificial intelligence innovation are complementary strategies. Using 75th-percentile thresholds to classify firms, we identify four distinct profiles: "AI Specialists" like Meta and Nvidia that lead in AI and ML patents but with few direct device filings; "Device Specialists" like Ubiquiti and Netgear that focus primarily on hardware and smart electronics; "Super Smart" firms like Microsoft and Honeywell that are in the top quartile of both smart device innovation and AI and machine learning patenting; and a long tail of firms below the top-quartile thresholds on both dimensions.

Taken together, these findings indicate that data capability has become increasingly characteristic of larger, more R&D-intensive, and more innovative public firms. Because

these differences persist after sector adjustment, they reflect real distinctions between data-capable and non-capable firms within similar industries, rather than merely across-sector variation in size and performance.

Appendix A.4 presents complementary evidence from early-stage private firms in PitchBook, where bringing a data-capable device to market predicts more successful exits in the first half of the sample.

3.1 Data-Capable Devices & Future Product Innovation

Table 2 examines whether firms that commercialize data-capable devices subsequently exhibit different product-introduction patterns. We identify "switcher" firms (grantees that had filed non-data products before their first data-device authorization) and require a minimum gap of three years between each firm's first non-data FCC filing and its first data-capable device filing, ensuring sufficient pre-adoption history to assess parallel trends in the event-study design.

To recover firms that adopted data-capable devices in the earliest cohorts (2007–2009), we extend the FCC filing history back to 2004; because connected devices with companion apps and cloud connectivity were essentially nonexistent before the smartphone era, all pre-2007 filings are classified as non-data devices. These criteria yield 209 switcher firms among US-listed grantees, with the narrow LLM classification determining both which devices count as data-capable and the timing of each firm's switch. Using a staggered adoption design, we compare each firm's post-switch filing activity (i.e., after it commercialized its first data-capable device) to its own pre-switch baseline, excluding the year of the switch itself to avoid mechanical effects. We control for the number of cumulative prior products and grantee age, both measured in the year prior to the switch. We report both standard Two-Way Fixed Effects and interaction-weighted estimates following Sun and Abraham (2021) to address potential bias from heterogeneous treatment timing.

Data-device adoption raises the likelihood of filing any new product in a given year by 13.8 percentage points (interaction-weighted estimate), against a baseline probability of 50.8% (Panel A), and adopters file approximately 1.2 more new products per firm-year against a baseline mean of about 2.3 (Panel B). Continued data-device filings account for a meaningful share of this increase (Panel C). Under the finest specification, which adds product-function-by-year fixed effects, the interaction-weighted estimates remain stable at about 0.8 additional products per year, while the extensive-margin filing-probability estimate stays positive but loses precision. Estimates for the full grantee panel of 1,943 switcher firms,

Table 2: Data-Capable Device Launches and Subsequent Product Filings among US Public Firms

	(1)	(2)	(3)	(4)
Panel A: Any New Product Filing				
Post-switch (TWFE)	0.293***	0.100***	0.218***	0.186***
	(0.029)	(0.032)	(0.046)	(0.045)
Post-switch (S&A)	—	0.138**	0.109	0.100
		(0.057)	(0.075)	(0.079)
Baseline Mean	0.508	0.508	0.508	0.508
Baseline SD	0.500	0.500	0.500	0.500
Observations	2,285	2,285	2,285	2,285
Panel B: N New Product Filings				
Post-switch (TWFE)	2.348***	0.827**	1.569***	1.286***
	(0.383)	(0.387)	(0.387)	(0.360)
Post-switch (S&A)	—	1.151**	0.989**	0.827*
		(0.460)	(0.473)	(0.483)
Baseline Mean	2.272	2.272	2.272	2.272
Baseline SD	4.810	4.810	4.810	4.810
Observations	2,285	2,285	2,285	2,285
Panel C: N New Data Device Filings				
Post-switch (TWFE)	1.417***	1.323***	1.706***	1.438***
	(0.316)	(0.325)	(0.417)	(0.323)
Post-switch (S&A)	—	0.643***	0.554*	0.660*
		(0.048)	(0.317)	(0.367)
Baseline Mean	0.658	0.658	0.658	0.658
Baseline SD	2.556	2.556	2.556	2.556
Observations	2,285	2,285	2,285	2,285
Grantee FE	No	Yes	Yes	Yes
Country × Year FE	No	No	Yes	Yes
Product Fn × Year FE	No	No	No	Yes

Notes: Sample restricted to switcher firms with at least three years of non-data FCC filings before their first data-device filing. Filing history extends to 2004 using archived FCC device records; all pre-2007 filings are treated as non-data devices, as connected products with companion apps were essentially nonexistent before the smartphone era. The adoption year (t=0) is excluded to avoid mechanical effects. Post-switch = 1 for years $t \geq 1$ after first data-device filing. S&A IW reports the average interaction-weighted post-treatment estimate (t=1 to t=4) following Sun and Abraham (2021), using the last-treated cohort as control. Controls: firm age at t= −1 and its square, IHS(prior products at t= −1). Standard errors clustered at the grantee level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

predominantly private and classified by keyword, are similar (not reported): adoption raises the probability of any new filing by 8.0 percentage points against a 38.0% baseline and adds 0.56 products per firm-year.[9] Given that the decision to develop a data-capable device is itself endogenous to a firm's product strategy, these estimates should be interpreted as descriptive associations rather than causal effects. Still, the findings are consistent with platform-style development: an initial data-enabled product anchors a stream of follow-on filings, raising both the probability and the volume of innovation within a firm's active development periods.

4 Data-Capability and Firm Outcomes

4.1 Background on GDPR

The EU's General Data Protection Regulation (GDPR), formally Regulation (EU) 2016/679, replaced the 1995 Data Protection Directive with a single, directly applicable EU-wide framework governing the processing of personal data. The regulation entered into force on May 24, 2016 and became applicable on May 25, 2018, following a two-year transition period. GDPR has explicit extraterritorial reach: it can apply to entities outside the EU when they offer goods or services to individuals in the EU or monitor their behavior there. GDPR is organized around the core processing principles of lawfulness, purpose limitation, data minimization, storage limitation, integrity/confidentiality, and accountability, and grants individuals enforceable rights (e.g., access, erasure, portability, objection).

GDPR is enforced by national supervisory authorities, coordinated through EU cooperation and consistency mechanisms (including the "one-stop-shop" for cross-border processing), with the European Data Protection Board (EDPB) playing a central role in promoting consistent application. The regulation raised compliance and enforcement stakes through (i) procedural obligations such as personal-data breach notification within 72 hours where feasible, and (ii) a two-tier administrative fine structure, including fines up to 20 million Euros or 4% of worldwide annual turnover for the most serious infringements.

GDPR is technology-neutral and therefore applies whenever personal data are processed through connected devices, companion apps, or cloud services. IoT products routinely generate data within GDPR's scope, including persistent device identifiers (IP addresses, RFID tags), location data, and telemetry, when linkable to an identifiable person.[10]

[9] In both samples the Sun and Abraham (2021) interaction-weighted estimates are smaller than the TWFE coefficients, consistent with upward bias in standard TWFE under staggered adoption.

[10] See: EU (2016); EuropeanCommission (2026); EDPB (2026)

4.2 Estimation and Results

We implement a triple differences-in-differences design from Equation 1 on a quarterly panel of public firms matched across Compustat and the FCC data to estimate the effect of the GDPR on EU-exposed firms with data-collecting devices. We focus on the quarters between 2011q1 and 2019q4 for this analysis. The event date is denoted by t^*, which here we assign as 2016q2, when the EU Parliament adopted the GDPR. The regulation was not implemented until 2018, but we look to include potential firm adjustments in anticipation of the GDPR during the phase-in period in the overall ATT of the GDPR.

$$\begin{aligned} Y_{it} = \beta_1[\mathbb{1}(t \geq t^*) \times \mathbb{1}_i(\text{EU-exposed}) \times D_i] + \beta_2[\mathbb{1}(t \geq t^*) \times \mathbb{1}_i(\text{EU-exposed})] \\ + \beta_3[\mathbb{1}(t \geq t^*) \times D_i] + \beta_4[\mathbb{1}_i(\text{EU-exposed}) \times D_i] \\ + \lambda_i + \eta_{qtr \times s(i)} + \epsilon_{it} \quad (1) \end{aligned}$$

We consider as outcomes Y_{it} the firms' quarterly market value, revenue, R&D spending, R&D intensity (as measured by R&D spending divided by pre-GDPR revenue), Tobin's Q, and Poisson counts of patents, citations, and citation-weighted patents for all patent filings on all technology classes and on AI/ML technology codes specifically. D_i is a measure of the firm's investment in data-capable devices. We consider several functional forms for D_i – a dummy for any data-capable devices, raw count of data-capable devices, and log data-capable devices – and indicate which is in use for a given set of results. Equation 1 includes firm fixed effects λ_i and quarter $\times$ 2-digit SIC industry $\eta_{qtr \times s(i)}$ fixed effects. We do not include fixed effects for the geography of firms' EU exposures because the granularity of EU exposure reporting is not consistent and often highly aggregated.

The main coefficient of interest here is β_1, which we interpret as the data-mediated impact of the GDPR on exposed firms to the extent that our core identification assumptions hold. These are that (A1) differences in trends between EU-exposed data-collectors and non-data collectors would have evolved similarly to the differences in trends between non-EU data-collectors and non-data collectors in a counterfactual without GDPR, (A2) the compositions of each of the four groups of firms (data-collects/not $\times$ EU-exposed/not) do not change differentially in the post period, and (A3) the GDPR does not affect outcomes for firms we classify as untreated (non-data-collecting or non-EU exposed).

While (A1) is not directly verifiable we estimate generally flat pre-trends in our event study plots, suggesting that groups did not necessarily have sustained differential evolutions

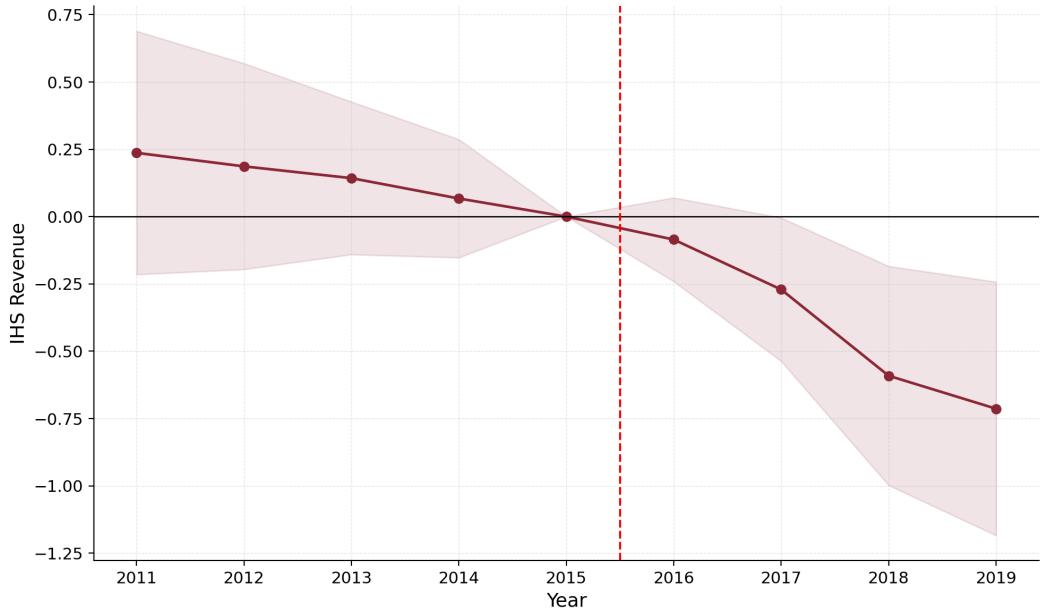
leading into GDPR. The primary threat to (A2) is that the GDPR drove differential compositional changes across groups, but Table A.5 compares the characteristics of firms that exit after GDPR to those who do not and finds no difference in data capability or EU exposure by exit status. Further, Tables A.8-A.8 show that the main results are robust to balancing the firm panel by including only firms alive throughout 2015q1 - 2019q4 (the year before GDPR to the end of the sample). (A3) is threatened if either our EU-exposure or data-collection measures are flawed or if non-EU or non-data-collecting firms are affected by GDPR in ways not controlled for by quarter × sector fixed effects. The EU-exposure measure is based on customer segment information described in 2.3 and we validate our data-collection measure in 2.2.

We partition the sample by 4-digit SIC industries into a set of consumer device producers (consumer electronics, household appliances, toys and sporting goods,...) and non-consumer device producers (industrial machinery, medical/surgical device production, aerospace and defense, and telecommunications equipment) as a falsification test of our consumer data mechanism: we plausibly do not expect restrictions on customer data transmission in the EU to affect non-consumer device manufacturers, so we separately identify the industry codes of consumer device producers and non-consumer device producers by excluding from the consumer set all device filers in industries like industrial machinery, medical/surgical device production, aerospace and defense, and telecommunications. Appendix Table A.3 presents the full list of SIC industry inclusions and exclusions used to create the consumer and non-consumer samples.

We demonstrate that this distinction is meaningful by estimating an alteration of Equation 1 among data-capable firms. We plot the coefficient on $consumer \times EU\text{-}exposed \times post\text{-}GDPR$ among all data-capable firms in Figure 4 for revenue. Consumer and non-consumer firms are selected so we interpret this as a purely descriptive result demonstrating that consumer device firms may have been affected by consumer data restrictions in a way non-consumer firms were not.

Table 4.2 presents β_1 from Equation 1 estimated for financial outcomes, where data capability is a binary indicator for whether the firm had data-capable device filings before the GDPR. Consumer firms are in Panel A and non-consumer firms are in Panel B. Treated consumer firms cut R&D spending by roughly 40 percent, while the estimates for market value, revenue, and Tobin's Q are negative but imprecise. The contraction also appears in R&D intensity (Column 4), where R&D is scaled by the firm's fixed 2015 revenue so that the denominator cannot respond to the shock; non-consumer intensity is unchanged. Non-consumer firms show no comparable financial response on any margin. Table 4.2 repeats this

Figure 4: Revenue of Consumer Relative to Non-Consumer Data-Capable Firms Around the GDPR



Note: The figure plots the yearly coefficients on the interaction of a consumer-industry indicator with EU exposure, estimated among data-capable firms only, with inverse-hyperbolic-sine revenue as the outcome and 2015 as the omitted year. The shaded region shows 95 percent confidence intervals from standard errors clustered by firm. Regressions include firm and quarter-by-industry fixed effects.

analysis using the count of data devices instead of a data-capability binary for the treatment. The dose margin sharpens the picture: each additional pre-GDPR data-capable device is associated with a 3.0 percent decline in R&D and a 2.3 percent decline in Tobin's Q, with imprecisely estimated effects on market value and revenue; non-consumer effects per device are small and, if anything, positive. Together these estimates indicate that GDPR's financial incidence on consumer data-device firms falls on investment and, increasing in the size of the device portfolio, on market valuation, while contemporaneous sales and profits are largely unaffected. Appendix Table A.4 decomposes these responses by the size of treated firms' pre-GDPR device portfolios: the Tobin's Q markdown is concentrated entirely in treated firms with above-median portfolios, firms holding only one or two data-capable devices are not repriced, and the decline in R&D intensity appears across both groups.

Table 4.2 presents β_1s on patent filing outcomes using the binary data-capability indicator for consumer firms in Panel A and non-consumer firms in Panel B. While patent filings of all types are unaffected, Columns 2 and 4 of Panel A demonstrate large, negative effects on AI/ML-related patent filing (down 31%) and citation-weighted AI/ML patent filing (down 89%) following the GDPR announcement. The difference in these effects suggests that the patent filings lost to this shock would have been high impact. In Panel B, non-consumer firms sharply increase both AI/ML and citation-weighted AI/ML patent filing. The estimated increases are very large in proportional terms, but they come off a small base: AI/ML patenting was rare among non-consumer device firms before 2016. Table A.7 repeats this analysis using the count of data devices instead of a data-capability binary. Patenting outcomes barely move with each additional device: overall filings tick up slightly, and AI/ML filings are essentially flat.

The estimated effects for all firms, but especially non-consumer firms, are massive. We re-estimate these effects in Table A.8 excluding the top 5 consumer patenters and the top 5 non-consumer patenters, respectively. The consumer AI/ML level effect attenuates, but the citation-weighted AI/ML decline persists, and if anything deepens, as does the non-consumer AI/ML increase. Thus the patent-level results are heavily influenced by, though not entirely attributable to, the most prolific AI/ML patent filers. These top 5 patent filers among consumer device producers are IBM, Alphabet, Microsoft, Amazon, and Ford; the top 5 among non-consumer device producers are Intel, Qualcomm, Apple, GE Aerospace, and HP.

Figure 5a shows an event study for patent filing for all consumer firms using the binary treatment. It demonstrates a clear trend break after the GDPR was announced in 2016. One pre-period coefficient, in 2014, deviates from zero. Event studies estimated on

Table 3: GDPR Exposure and the Financial Outcomes of Data-Capable Firms

Panel A: Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Data	-0.386	-0.261	-0.508**	-0.352**	-0.055
	(0.259)	(0.182)	(0.210)	(0.166)	(0.229)
Dep. var. mean	8.42	6.26	3.50	0.24	1.34
Dep. var. SD	2.66	2.75	2.45	0.62	0.89
Observations	5,176	5,342	3,582	3,473	5,170
Firms	164	164	147	139	164
Adj. R-sq	0.963	0.981	0.977	0.642	0.779
Panel B: Non-Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Data	0.319	0.230	0.064	0.020	-0.035
	(0.216)	(0.160)	(0.142)	(0.027)	(0.137)
Dep. var. mean	8.29	6.39	3.71	0.13	1.18
Dep. var. SD	2.47	2.36	2.07	0.15	0.59
Observations	5,891	6,054	4,898	4,831	5,891
Firms	184	184	172	164	184
Adj. R-sq	0.969	0.981	0.970	0.835	0.742

Notes: The table reports the coefficient on the triple interaction of EU exposure, the post-GDPR period, and the firm's data-capability treatment from Equation 1, estimated on a quarterly panel of consumer device firms (Panel A) and non-consumer device firms (Panel B). The treatment is an indicator for any pre-GDPR data-capable device, and the sample is an unbalanced panel of firms that enter by the first quarter of 2015. All regressions include firm and quarter-by-industry fixed effects, and standard errors are clustered by firm. R&D intensity is quarterly R&D spending divided by the firm's average quarterly revenue in 2015, so that the denominator is fixed before the regulation; ratios above ten are dropped and the remainder is winsorized at the 1st and 99th percentiles within each panel. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: GDPR Exposure and Financial Outcomes per Additional Data-Capable Device

Panel A: Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Count	-0.030	-0.011	-0.030***	-0.017	-0.023*
	(0.019)	(0.010)	(0.011)	(0.010)	(0.013)
Dep. var. mean	8.42	6.26	3.50	0.24	1.34
Dep. var. SD	2.66	2.75	2.45	0.62	0.89
Observations	5,176	5,342	3,582	3,473	5,170
Firms	164	164	147	139	164
Adj. R-sq	0.963	0.981	0.976	0.638	0.779
Panel B: Non-Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Count	0.026***	0.014**	0.011*	0.001	-0.003
	(0.008)	(0.006)	(0.006)	(0.001)	(0.005)
Dep. var. mean	8.29	6.39	3.71	0.13	1.18
Dep. var. SD	2.47	2.36	2.07	0.15	0.59
Observations	5,891	6,054	4,898	4,831	5,891
Firms	184	184	172	164	184
Adj. R-sq	0.969	0.982	0.970	0.835	0.743

Notes: The table reports the coefficient on the triple interaction of EU exposure, the post-GDPR period, and the firm's data-capability treatment from Equation 1, estimated on a quarterly panel of consumer device firms (Panel A) and non-consumer device firms (Panel B). The treatment is the count of the firm's pre-GDPR data-capable devices, and the sample is an unbalanced panel of firms that enter by the first quarter of 2015. All regressions include firm and quarter-by-industry fixed effects, and standard errors are clustered by firm. R&D intensity is quarterly R&D spending divided by the firm's average quarterly revenue in 2015, so that the denominator is fixed before the regulation; ratios above ten are dropped and the remainder is winsorized at the 1st and 99th percentiles within each panel. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: GDPR Exposure and the Patenting of Data-Capable Firms

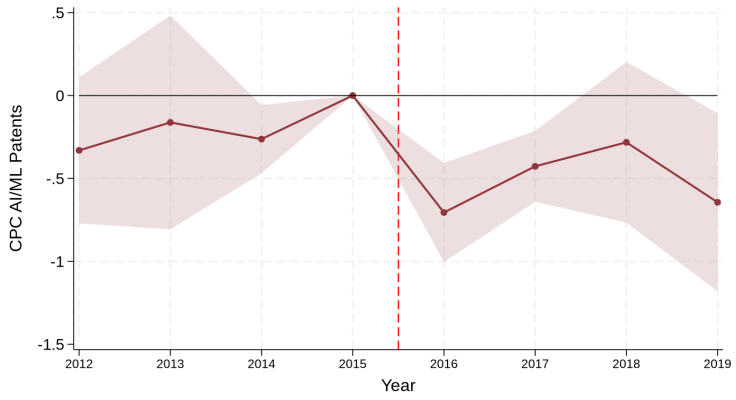
Panel A: Consumer					
	(1)	(2)	(3)	(4)	(5)
	Patents	AI/ML	CW Pat	CW AI/ML	Val Pat ($)
EU × Post × Data	0.024	-0.372***	-0.219	-2.240***	-0.287
	(0.301)	(0.115)	(0.456)	(0.220)	(0.247)
Dep. var. mean	219	27	407	1595	4181
Dep. var. SD	837	110	2450	4052	13938
Observations	949	334	933	275	949
Firms	110	45	108	37	110
Panel B: Non-Consumer					
	(1)	(2)	(3)	(4)	(5)
	Patents	AI/ML	CW Pat	CW AI/ML	Val Pat ($)
EU × Post × Data	-0.325	3.568***	-0.080	3.687***	0.491**
	(0.460)	(1.118)	(0.339)	(1.193)	(0.234)
Dep. var. mean	143	5	98	252	2109
Dep. var. SD	403	16	285	743	9895
Observations	1,245	465	1,202	366	1,245
Firms	144	56	138	43	144

Notes: The table reports Poisson pseudo-maximum-likelihood estimates of the triple-difference coefficient on annual patenting outcomes. The sample is an unbalanced panel of firms that enter by the first quarter of 2015. All regressions include firm and year-by-industry fixed effects, and standard errors are clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

IHS-transformed outcomes do not exhibit pre-trends, suggesting that the pre-period Poisson coefficients are potentially driven by the sensitivity of count models to large outliers in thin cells rather than by systematic and multi-period differential trends. Our ongoing work to expand the sample aims to improve precision for these analyses. Still, the declining post-2016 pattern in AI and machine learning patent applications, as well as the ATT results presented in the tables above, are consistent with the interpretation that EU-exposed data-capable firms saw meaningful innovation slowdowns after GDPR's restrictions on personal data use were announced.

Figure 5: Patent Filing by Consumer Device Firms Around the GDPR

(a) AI/ML Patents



To further examine the changes in AI-related innovation around GDPR, we use the AIPD dataset, the United States Patent and Trademark Office's Artificial Intelligence Patent Dataset (Giczy et al., 2022; United States Patent and Trademark Office, 2023). Rather than identifying AI through patent classification codes, the AIPD classifies every US patent using machine-learning models trained on patent text, and it decomposes AI into eight component technologies: machine learning, evolutionary computation, natural language processing, speech recognition, computer vision, planning and control, knowledge processing, and AI hardware. This provides an independent, text-based measure of AI activity and, more importantly, lets us ask which kinds of AI innovation respond to the regulation. Table 6 reports the triple-difference estimates for consumer firms, pooling all AI components in the first row and estimating the same specification separately for each component below; Appendix Table A.11 reports the non-consumer sample.

The composition of the response sharpens the mechanism. The contraction concentrates precisely in the components that consume personal sensory data from consumer de-

vices: computer vision patenting falls by roughly half, and speech recognition declines as well. These are the technologies trained on the images, audio, and behavioral streams that consumer devices collect and that GDPR regulates. Components that operate on abstract methods or non-personal corpora, such as core machine learning, evolutionary computation, and planning and control, rise instead, so total AI patenting under this alternative measure is roughly unchanged. Among non-consumer firms, the vision and speech declines are absent. With a measure of AI constructed entirely independently of our patent-code approach, the same conclusion emerges: after GDPR, the innovation that slows is the innovation that is most reliant on personal data, while treated firms redirect their AI efforts toward applications the regulation does not touch.

Table 6: GDPR and the Composition of AI Patenting: Consumer Firms

Patent category	EU × Post × Data	SE	Firms	Observations
Any AI (all components pooled)	0.326	(0.218)	81	704
Machine learning	0.585***	(0.093)	44	364
Evolutionary computation	1.044***	(0.116)	37	302
Natural language processing	0.812***	(0.141)	42	326
Speech recognition	-0.220*	(0.115)	41	321
Computer vision	-0.689***	(0.194)	63	539
Planning and control	0.762***	(0.184)	70	606
Knowledge processing	0.766***	(0.144)	54	454
AI hardware	0.188	(0.213)	50	413

Notes: The table reports Poisson pseudo-maximum-likelihood estimates of the triple-difference coefficient from Equation 1 on annual counts of AI patents, where patents are classified into AI and its component technologies by the USPTO Artificial Intelligence Patent Dataset's text-based models rather than by CPC codes. The first row pools all AI components; the remaining rows estimate the same specification separately for each component. The sample is the unbalanced panel of consumer device firms that enter by the first quarter of 2015. All regressions include firm and year-by-industry fixed effects, and standard errors are clustered by firm. Poisson estimation retains only firms with variation in the outcome, so cluster counts vary across rows. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5 Discussion and Conclusion

This paper develops a novel, product-level measure of data capital by applying natural language processing to the user-manual texts of wireless devices authorized by the FCC. We document that data-capable connected devices ("smart" products that collect and transmit user data via companion apps and cloud services) are increasingly characteristic of large, R&D-intensive, and innovative public firms, and that prior data-device commercialization is associated with more prolific subsequent product introductions.

Our central empirical contribution exploits the GDPR as a shock restricting firms' ability to collect and process personal data from European customers. Using a triple-difference design that compares EU-exposed data-capable consumer firms to (i) EU-exposed non-data firms, (ii) non-EU data firms, and (iii) non-consumer data firms, we find that GDPR adoption is associated with a roughly 40% cut in R&D spending, a 31% decline in AI/ML patent filings, and an 89% decline in citation-weighted AI/ML patents among treated consumer firms. The investment response also appears in R&D intensity measured against pre-GDPR revenue, and it scales with exposure: each additional data-capable device in a firm's pre-GDPR portfolio is associated with a further 3% decline in R&D. The composition of the patenting response points to the same mechanism. Under the USPTO's independent, text-based AI patent classification, the contraction concentrates in computer vision and speech recognition, the components trained on the images, audio, and behavioral streams that consumer devices collect, while method-oriented AI components rise.

Financial markets appear to recognize the loss: Tobin's Q falls for treated consumer firms with substantive data-device portfolios, consistent with a durable markdown of the expected rents from personal data, even as contemporaneous sales and profits hold steady. Non-consumer firms, whose connected devices serve industrial rather than personal-data applications, show none of these declines, and if anything exhibit increased patenting, consistent with GDPR selectively disrupting innovation that depends on personal consumer data.

These findings have implications for how we understand data as a form of intangible capital. Our findings about innovation suggest that first-party data functions as a productive, depreciable asset: it is accumulated through deliberate investment in data-capable hardware, it is durable and reusable across production cycles, and it complements R&D in producing innovative output. When regulation restricts the collection and reuse of this capital stock through GDPR's purpose-limitation, data-minimization, and erasure requirements, the returns to complementary R&D investment fall. Firms respond both by investing less and by redirecting their innovative effort away from applications that depend on personal

data, such as computer vision and speech recognition, toward uses the regulation does not reach.

We see room for improvement in two areas of ongoing work. The first is statistical power. Our triple-difference design relies on variation from roughly 45 firms that are simultaneously EU-exposed, data-capable, and consumer-facing: this is a small group, even if the patent results survive excluding the most prolific patenters. The R&D contraction is precisely estimated, but the remaining financial outcomes are not: with 45 treated firms, we may not have enough variation to detect effects on noisy quarterly outcomes like market value or revenue. We are working to expand and improve the panel by tightening the FCC-to-Compustat matching to bring in additional firms, and incorporate international firm data that would let us exploit richer cross-country variation in GDPR exposure.

The second is that several of our variables are more blunt than would be ideal to capture real-world phenomena. EU exposure comes from geographic segment disclosures, which firms report under a vague materiality standard, so some firms we classify as unexposed may in fact sell to European customers, and some we classify as exposed may have minimal EU revenue. The fact that results survive when we swap in continuous EU revenue shares or broader international-exposure indicators is encouraging, but we are also building more granular measures from subsidiary-level CapitalIQ data. The consumer/non-consumer split is similarly rough: it relies on SIC codes, which do not always map cleanly onto whether a firm's connected devices end up in households or on factory floors. Our ongoing work also aims to build device- and firm-level customer-composition measures to do better here. Our data capability indicator itself is a simple flag for whether a firm sells any connected device; it does not distinguish a company with a single Bluetooth speaker from one with an ecosystem of cloud-connected products and a broad user base collecting terabytes of data. Thus, our ongoing work aims to build richer measures that can help capture the intensity of data collection more directly.

Notwithstanding these limitations, the paper provides new evidence that data from connected devices acts as a productive intangible asset with material implications for innovation. Our results suggest that the costs of restricting its collection arrive not only as less innovation, but as different innovation. As the IoT continues to grow and privacy regulation spreads beyond Europe, the tension between protecting consumer data and preserving the innovation incentives that data creates will only become more pressing.

References

Aridor, G., Che, Y.-K., & Salz, T. (2023). The effect of privacy regulation on the data industry: empirical evidence from GDPR. *RAND Journal of Economics*, *54*(4), 695–730. Retrieved 2026-01-28, from https://ideas.repec.org//a/bla/randje/v54y2023i4p695-730.html (Publisher: RAND Corporation)

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, *151*(C). Retrieved 2026-01-29, from https://ideas.repec.org//a/eee/jfinec/v151y2024ics0304405x2300185x.html (Publisher: Elsevier)

Bergeaud, A., Jaffe, A. B., & Papanikolaou, D. (2025, May). *Natural Language Processing and Innovation Research* [Working Paper]. National Bureau of Economic Research. Retrieved 2026-01-27, from https://www.nber.org/papers/w33821 doi: 10.3386/w33821

Blind, K., Niebel, C., & Rammer, C. (2024). The impact of the EU general data protection regulation on product innovation. *Industry and Innovation*, *31*(3), 337–357.

Board, F. A. S. (2026). *Accounting standards codification.* Retrieved from https://asc.fasb.org/

Boroomand, F., Leiponen, A., & Vasudeva, G. (2022). Does the Market Value Attention to Data Privacy? Evidence from U.S.-listed Firms under the GDPR.

BusinessWire.com. (2021). *Data Creation and Replication Will Grow at a Faster Rate Than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts.* Retrieved 2026-02-01, from https://www.businesswire.com/news/home/20210324005175/en/Data-Creation-and-Replication-Will-Grow-at-a-Faster-Rate-Than-Installed-Storage-Capacity-According-to-the-IDC-Global-DataSphere-and-StorageSphere-Forecasts

Chung, K. H., & Pruitt, S. W. (1994). A simple approximation of tobin's q. *Financial Management*, *23*(3), 70–74. Retrieved 2026-02-01, from http://www.jstor.org/stable/3665623

Crouzet, N., & Eberly, J. (2023). Rents and Intangible Capital: A Q+ Framework. *The Journal of Finance*, *78*(4), 1873–1916. Retrieved 2026-02-02, from https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13231 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13231) doi: 10.1111/jofi.13231

Crouzet, N., Eberly, J. C., Eisfeldt, A. L., & Papanikolaou, D. (2022, August). The Economics of Intangible Capital. *Journal of Economic Perspectives*, *36*(3), 29–52. Retrieved 2026-02-02, from https://www.aeaweb.org/articles?id=10.1257%2Fjep.36

.3.29 doi: 10.1257/jep.36.3.29

de Cornière, A., & Taylor, G. (2024, September). Data-Driven Mergers. *Management Science*, *70*(9), 6473–6482. Retrieved 2026-01-27, from https://pubsonline.informs.org/doi/10.1287/mnsc.2023.04104 (Publisher: INFORMS) doi: 10.1287/mnsc.2023.04104

Demirer, M., Jiménez Hernández, D. J., Li, D., & Peng, S. (2024, February). *Data, Privacy Laws and Firm Production: Evidence from the GDPR* [Working Paper]. National Bureau of Economic Research. Retrieved 2026-01-28, from https://www.nber.org/papers/w32146 doi: 10.3386/w32146

EDPB. (2026). *GDPR Cooperation and Enforcement | European Data Protection Board.* Retrieved 2026-01-29, from https://www.edpb.europa.eu/our-work-tools/support-cooperation-and-enforcement/gdpr-cooperation-and-enforcement_en?utm_source=chatgpt.com

EU, R. (2016, April). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).* Retrieved 2026-01-29, from http://data.europa.eu/eli/reg/2016/679/oj (Legislative Body: EP, CONSIL)

EuropeanCommission. (2026). *Data protection explained - European Commission.* Retrieved 2026-01-29, from https://commission.europa.eu/law/law-topic/data-protection/data-protection-explained_en

Ewens, M., Peters, R. H., & Wang, S. (2025, January). Measuring Intangible Capital with Market Prices. *Management Science*, *71*(1), 407–427. Retrieved 2026-01-28, from https://pubsonline.informs.org/doi/10.1287/mnsc.2021.02058 (Publisher: INFORMS) doi: 10.1287/mnsc.2021.02058

Farboodi, M., Mihet, R., Philippon, T., & Veldkamp, L. (2019, May). Big Data and Firm Dynamics. *AEA Papers and Proceedings*, *109*, 38–42. Retrieved 2026-01-27, from https://www.aeaweb.org/articles?id=10.1257/pandp.20191001 doi: 10.1257/pandp.20191001

Farboodi, M., & Veldkamp, L. (2025). *FV_datagrowth.pdf.* Retrieved 2026-01-28, from https://drive.google.com/file/u/1/d/1dXcOXTwdKP40KPk2agxBUUxSdSPu-vYk/view?usp=sharing&usp=embed_facebook

Giczy, A. V., Pairolero, N. A., & Toole, A. A. (2022). Identifying artificial intelligence (AI) invention: a novel AI patent dataset. *The Journal of Technology Transfer*, *47*, 476–505.

Giorgio, M., Benedikt Frey, C., & Andres, P. (2024). *Redirecting AI: Privacy regulation and the future of artificial intelligence* (VoxEU). Oxford Martin School. (Available at https://cepr.org/voxeu/columns/redirecting-ai-privacy-regulation-and-future-artificial-intelligence)

Goldberg, S. G., Johnson, G. A., & Shriver, S. K. (2024, February). Regulating Privacy Online: An Economic Evaluation of the GDPR. *American Economic Journal: Economic Policy*, *16*(1), 325–358. Retrieved 2026-01-28, from https://www.aeaweb.org/articles?id=10.1257%2Fpol.20210309&utm_source=chatgpt.com doi: 10.1257/pol.20210309

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2005). Market value and patent citations. *The RAND Journal of Economics*, *36*(1), 16–38.

IOT-Analytics.com. (2025). *Number of connected IoT devices growing 14% to 21.1 billion.* Retrieved 2026-02-01, from https://iot-analytics.com/number-connected-iot-devices/

Jones, C. I., & Tonetti, C. (2020, September). Nonrivalry and the Economics of Data. *American Economic Review*, *110*(9), 2819–2858. Retrieved 2026-01-27, from https://www.aeaweb.org/articles?id=10.1257/aer.20191330 doi: 10.1257/aer.20191330

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017, May). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics*, *132*(2), 665–712. Retrieved 2026-01-27, from https://doi.org/10.1093/qje/qjw040 doi: 10.1093/qje/qjw040

Liu, P.-Y. (2024). Value of Data to U.S. Public Firms: Evidence from GDPR.

Quan, J. (2024). *Junjun_quan_value_of_data.pdf.* Working Paper. Retrieved 2026-01-29, from https://www.dropbox.com/scl/fi/54epblhmp377v12h00itc/Junjun_Quan_Job_Market_Paper.pdf?dl=0&rlkey=zwzilgvfdhsrg665hrpgkt55t

Sun, L., & Abraham, S. (2021, December). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. doi: 10.1016/j.jeconom.2020.09.006

Szkutak, R. (2025, January). *VCs say AI companies need proprietary data to stand out from the pack.* Retrieved 2026-01-28, from https://techcrunch.com/2025/01/10/vcs-say-ai-companies-need-proprietary-data-to-stand-out-from-the-pack/

United States Patent and Trademark Office. (2023). *Artificial intelligence patent dataset (AIPD), 2023 update.* https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset.

A Appendix

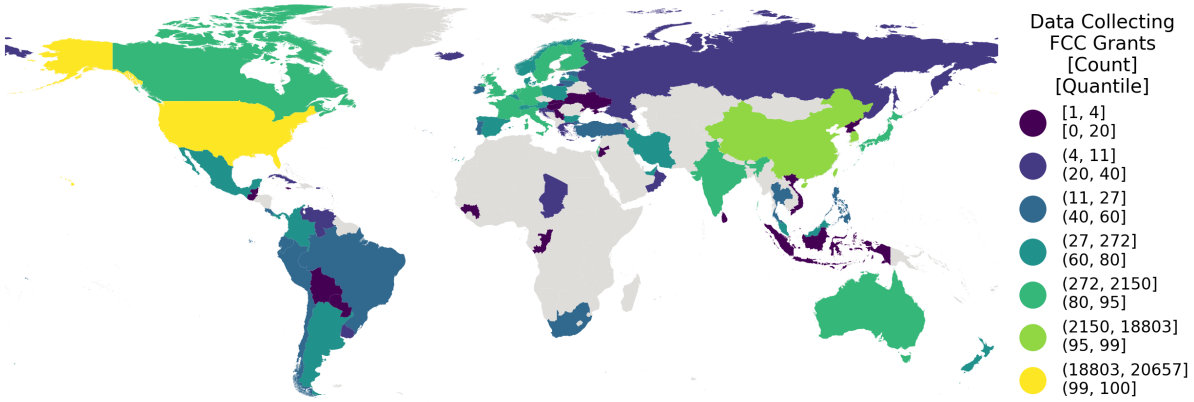
Table A.1: Mean keyword frequency by category: data-device vs. non-data-device deals in SEC acquisition filings

	Data (n=129)	No Data (n=46)	Diff	p-value
Engagement / Retention	0.39	0.02	+0.37	0.004***
IoT / Connected Devices	0.61	0.59	+0.03	0.953
Data Collection	0.16	0.20	−0.03	0.865
Monetization	0.06	0.00	+0.06	0.207
Advertising	0.50	0.50	−0.00	0.993
Unit Economics	0.71	0.11	+0.60	0.334
Personalization	0.07	0.13	−0.06	0.503

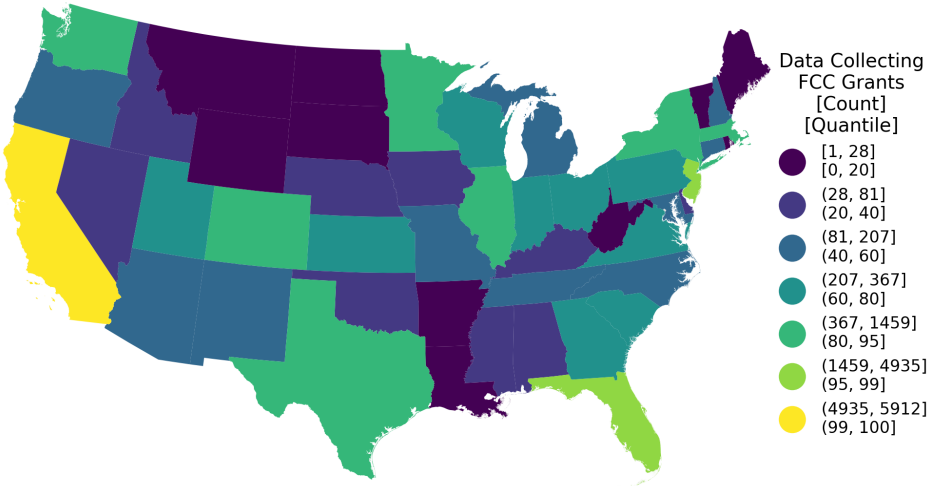
Notes: All deals (N=175). Mean keyword mentions per deal. Welch's t-test used for difference in means. "Data" refers to deals where the matched firm is classified as a data-device company; "No Data" refers to deals where the matched firm is not. Sample restricted to high-quality FCC-SEC matches (score ≥ 0.9).
*** p<0.01, ** p<0.05, * p<0.10

Figure A.1: Approved FCC filings by grantee country and state headquarters, 2007-2025

(a) Approved FCC filings by country



(b) Approved FCC filings by state



Note: Quantiles and corresponding counts are computed based on entities with at least one approved FCC filing for a data-collecting device.

Figure A.2: Sector-adjusted comparisons across data-capable versus non-capable public firms (2007-2023)

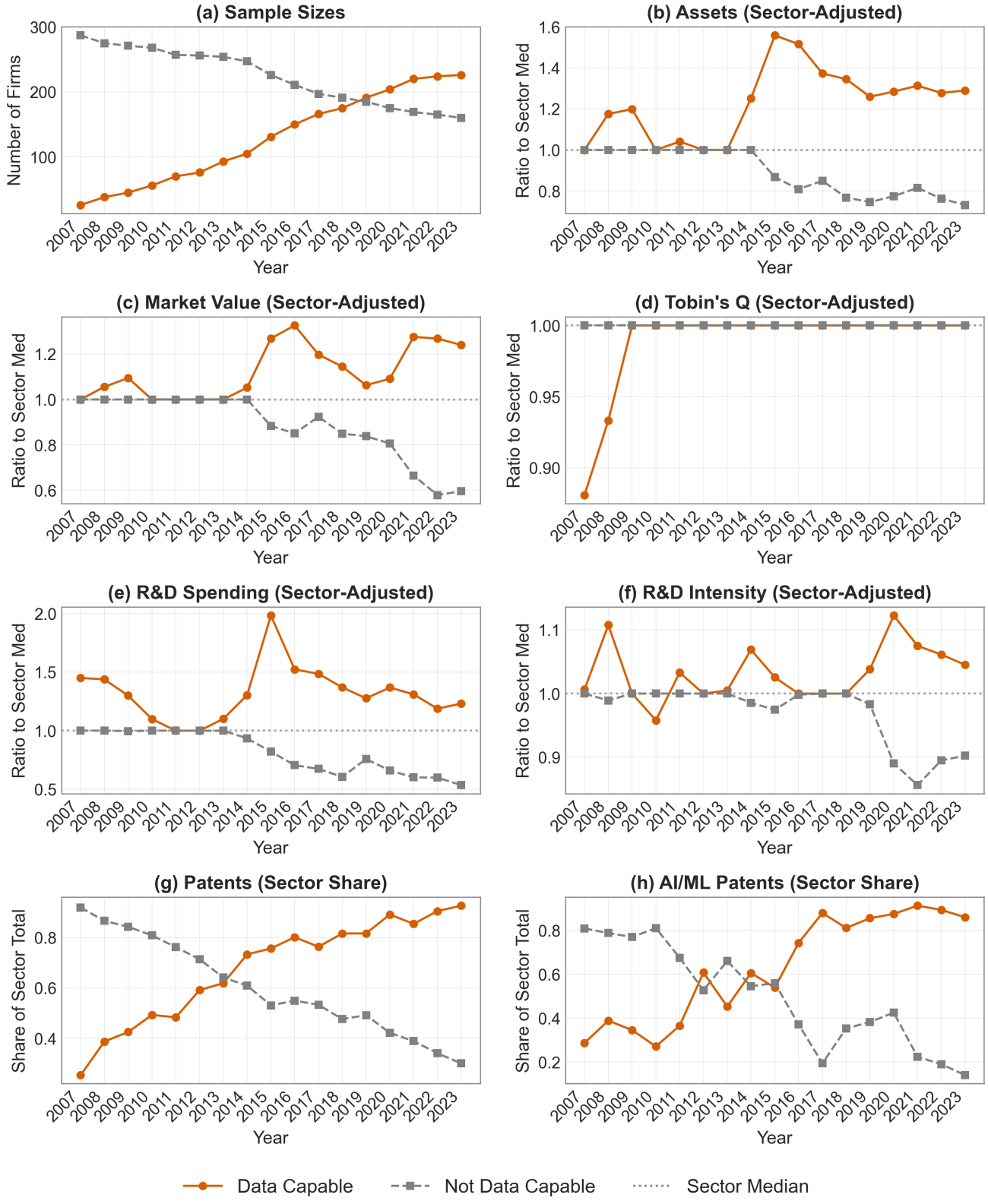
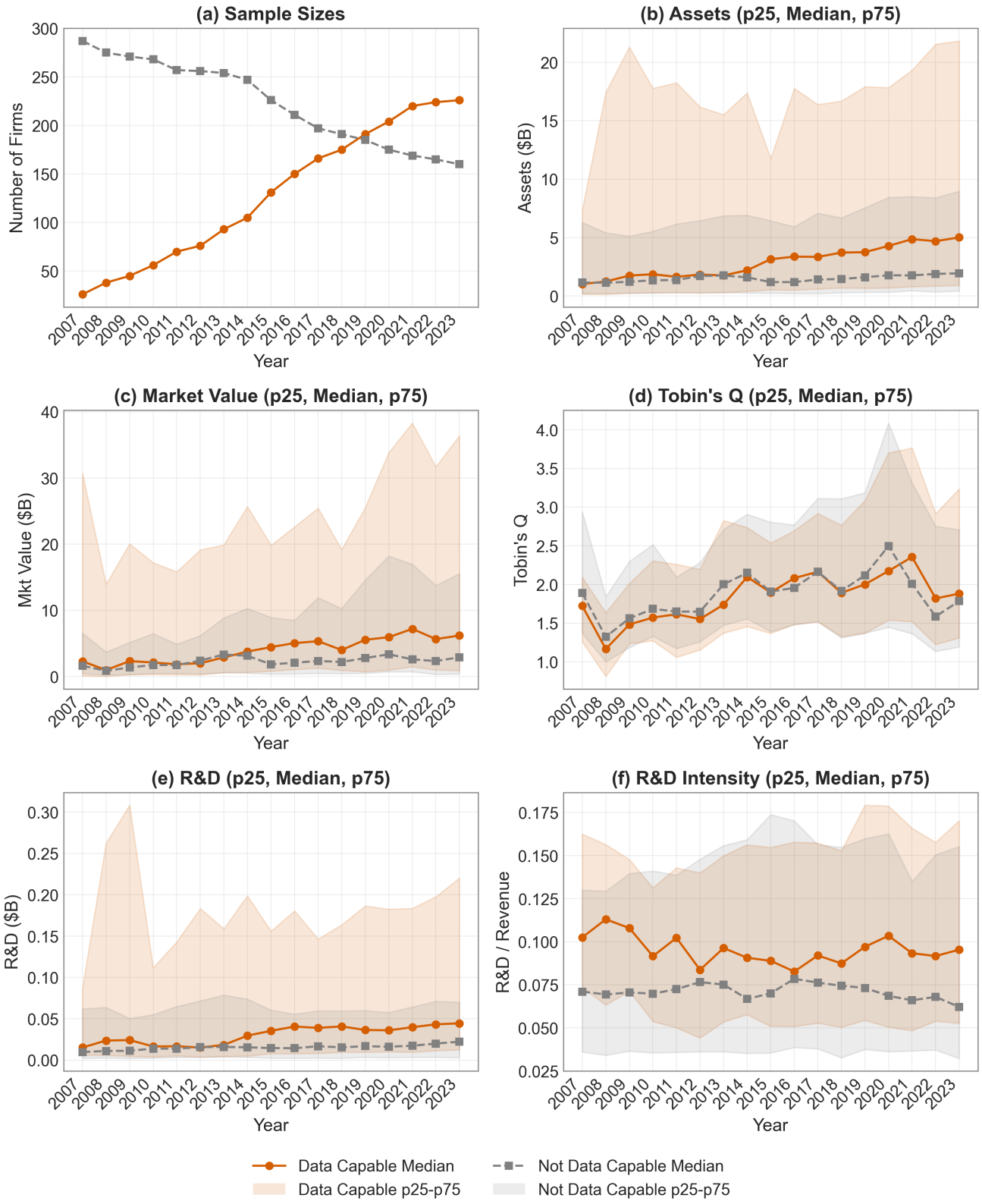


Figure A.3: Comparing data-capable versus non-capable public firms on the basis of medians, and 25th and 75th percentiles, (2007-2023)



A.1 Details of the LLM Classification of User Manuals

Extraction. We process each manual with Anthropic's Claude Opus 4.8 through the provider's batch interface. The model receives the full extracted manual text (truncated at 240,000 characters, which binds for very few documents) together with a fixed instruction set, and it must respond through a single structured tool call that records the device profile. Because the response format is constrained by a schema, the output requires no post-hoc parsing: every manual yields the same set of typed fields. The core fields are

1. `app_present`, `app_required`, `app_optional`: whether a mobile companion app (a phone or tablet app from the App Store or Google Play, or a named brand app such as the "Ring app") exists, and whether device setup or a core function needs it. Desktop and PC software explicitly does not count here.
2. `account_present`: whether operating the device involves creating or signing into a user account.
3. `cloud_present`: whether the device sends data to or retrieves data from a cloud service.
4. `desktop_software`: whether the manual names companion software that runs on a PC (a Windows configuration utility, a vendor desktop application).
5. `pairing_language`: whether the manual describes an explicit pairing or network-setup flow.

For every field set to one, the model must return a verbatim quotation from the manual as supporting evidence, and the quotations are stored alongside the flags in the classification file. The instruction set imposes three rules that matter for measurement. First, the model may use only the manual text; it may not draw on outside knowledge of the brand or product line. Second, when the text is silent the field stays at its default of zero, so ambiguity is resolved conservatively rather than by inference. Third, documents that are not user manuals (test reports, compliance letters) receive no positive flags. A device is data-capable under the narrow measure if `app_present`, `account_present`, or `cloud_present` equals one, and under the broad measure if any of the five fields equals one. The marginal cost of classification is approximately \$0.05 per manual, so the method scales to regulatory text corpora far larger than ours.

Checks. A second language model (Claude Sonnet 4.6) classifying a stratified sample of 98 manuals under the same instructions agrees with the primary model on 92 to 100 percent of

devices, depending on the indicator. An independent audit by a more capable model (Claude Fable 5) of 500 manuals, oversampling devices on which the LLM and keyword codings disagree, confirms 95 percent of the classification's positive calls, agrees on all unanimously coded control devices, and finds that residual errors under-count rather than over-count data capability. Hand adjudication of 150 keyword/LLM disagreements attributes most of them to two keyword error categories: PC-configuration utilities read as app connectivity, and matches inside compliance documents mislabeled as manuals. The pipeline is deterministic; re-running the extraction and dataset construction from scratch reproduces the classification file byte for byte. Substituting the keyword coding into the identical downstream pipeline reproduces the original keyword-based estimates exactly, so differences between results under the two codings reflect the measure rather than the pipeline.

A.2 Details of NLP Algorithm to Identify Data-Capable Devices

1. Document Filtering: In some instances, some FCC application documents like test reports or cover letters that are not user manuals are saved under that label. So our first step was to exclude such documents by searching for terms characteristic of test reports and compliance documentation (e.g., "test report," "EMC test," "SAR test," "RF exposure evaluation," "label location"). Documents shorter than 500 characters are also excluded.

2. Primary Keyword Detection: A device is classified as having a companion app if its user manual contains any single match from a set of high-confidence patterns. These include: (i) explicit app-type mentions such as "mobile app," "smartphone app," "companion app," "iOS app," or "Android app"; (ii) app installation references such as "download the app" or "install the app"; (iii) app store references including "App Store," "Google Play," or "Play Store"; (iv) brand-specific app names for major IoT platforms and device manufacturers (e.g., Ring, Wyze, Arlo, Nest, Eufy, Tuya, SmartLife, Kasa, Alexa, Google Home); and (v) direct app references such as "our app," "use the app to," or "in-app settings."

3. Secondary Keyword Detection: If no primary keywords are found, a device is classified as data-capable if two or more secondary indicators are present. Secondary keywords capture: (i) smart home platform references (SmartLife, Tuya, eWelink, SmartThings); (ii) generic app functionality phrases ("through the app," "in the app," "app control"); (iii) WiFi/Bluetooth pairing and network setup language; (iv) remote access patterns ("remote view," "view anywhere," "monitor remotely," "live stream"); (v) push noti-

fication and alert patterns; (vi) account and cloud patterns ("create account," "sign up," "cloud storage," "cloud recording"); and (vii) QR code setup references.

4. Other Data Collection Indicators: Independent of app detection, we flag devices whose manuals reference explicit data transmission or storage activities, including "data sync," "upload data," "usage data," "collect data," "cloud storage," "store to server," or "transmit data."

A device is classified as "data-capable" if it satisfies either the companion-app criteria (items #2–3) or the other-data-collection criteria (#4). Using this approach, more than a third of devices in our sample are classified as data-capable, with the rate rising over time consistent with the growth of the Internet of Things.

A.3 Details of the Matching Algorithm

1. Standardize firm details (name, address, contact information) across FCC grantees, CapitalIQ, and PitchBook.

2. Exact match FCC grantees to CapitalIQ or PitchBook using firm legal name.

3. Exact match FCC grantees to CapitalIQ or PitchBook using contact information and selecting for the match with the highest name similarity.

4. Exact match on the soundex encoding of each firm's name, the first word of each firm's name, the reported state and city of each firm, or the first three letters of the firm's name, keeping matches where firm names without legal identifiers display over 50 percent similarity.

5. Across the resulting batch of candidates, keep those with an exact match on name and contact information, then keep candidate matches based on the following prioritized sequence of criteria:

   (a) over 95 percent similarity in firm legal names.

   (b) over 75 percent similarity in firm legal names and street addresses.

   (c) over 95 percent similarity of one candidate firm's name against a substring of the other candidate firm's name.

6. For firms mapped to CapitalIQ, map the corresponding CapitalIQ firm to Compustat, or the parent firm's Compustat record if an operating subsidiary, using GVKeys.

A.4 Data Capability among Early-Stage Firms

Beyond looking at public firms, we also study the characteristics and outcomes of data-capable early-stage startups using PitchBook's deal data. The estimates in this section are based on the complementary rule-based text measure described in Appendix A.2; re-estimation under the LLM-based manual-read classification is in progress.

We estimate a discrete hazard model from Equation 2 for several outcomes for US PitchBook firms with respect to device type filings. Let $Y_{it} \in \{0, 1\}$ where Y represents an outcome from {bankruptcy, buyout, merger/acquisition, IPO} and $Y_{it} = 1 \implies Y_{i,t+j}$ are censored for all $j > 0$. N_{it}^X is the number of devices of type X the firm i has filed by t and N_{it} is the firm's count of devices of all types, $X \in$ {data-collecting, wifi-enabled, has companion app}.

$$\Pr(Y_{it} = 1 \mid active) = \beta_1 \mathbb{1}\{N_{it}^X > 0\} + \beta_2 \log(1 + N_{it}) + C\gamma + \lambda_{s(i)\times t} + \lambda_{g(i)} + \epsilon_{it} \quad (2)$$

The control term C contains firm age bins, indicators for the type of the firm's most recent deal (private equity, venture, corporate debt, etc.), and indicators for the most recent firm status observed as of t: profitability, revenue-generating, or trials/development. The λ terms are fixed effects for sector-time $s(i) \times t$ and state $g(i)$. The sample is all firms who ever file a device of type X in the time-frame with years after each firm's first outcome Y event excluded.

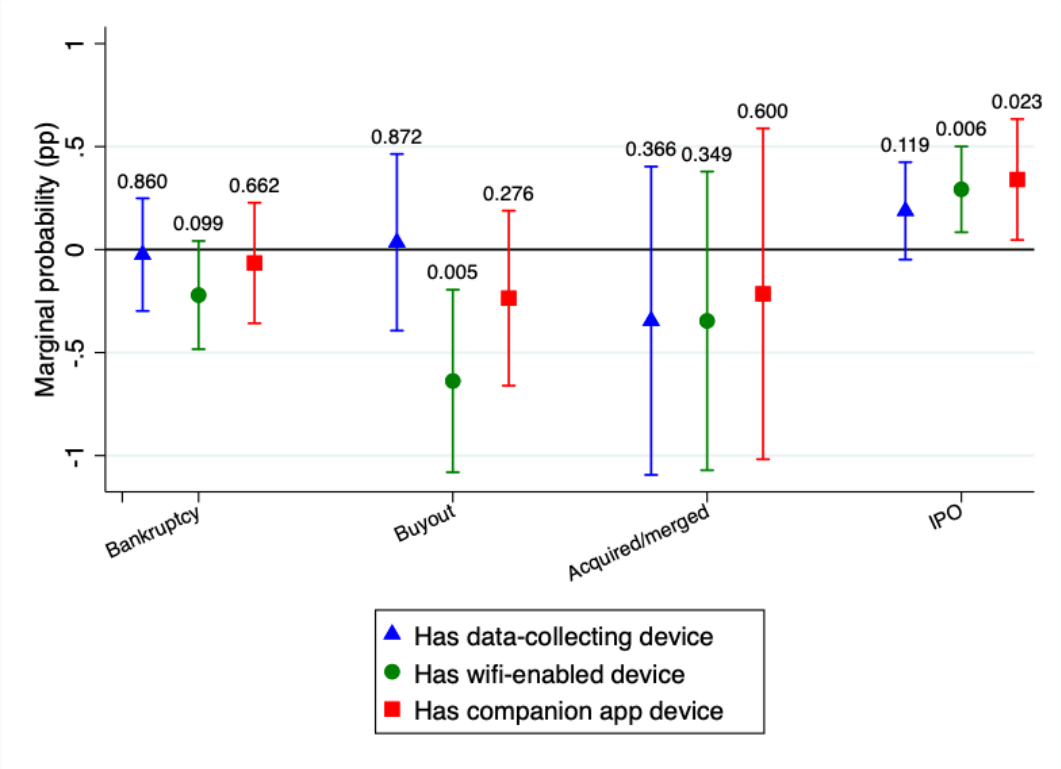
Figure A.4 plots β_1 for each outcome from Equation 2 from 2007-2013 in Panel A and from 2014-2019 in Panel B. We split the sample in 2014 to consider the trajectories of data-collecting firms in two periods that appear distinct; as Figure 3 demonstrates, 2014 is the first year in which the median data-collector had more assets, a higher valuation, and spent more on R&D than the median non-data-collector.

Figure A.4 shows that, conditional on age, deal histories, profitability, and fixed effects, both firms with wifi-enabled devices and companion app devices were 0.2-0.3 percentage points more likely to IPO than firms without, and firms with wifi-enabled devices were more than half a percentage point less likely to be bought out.

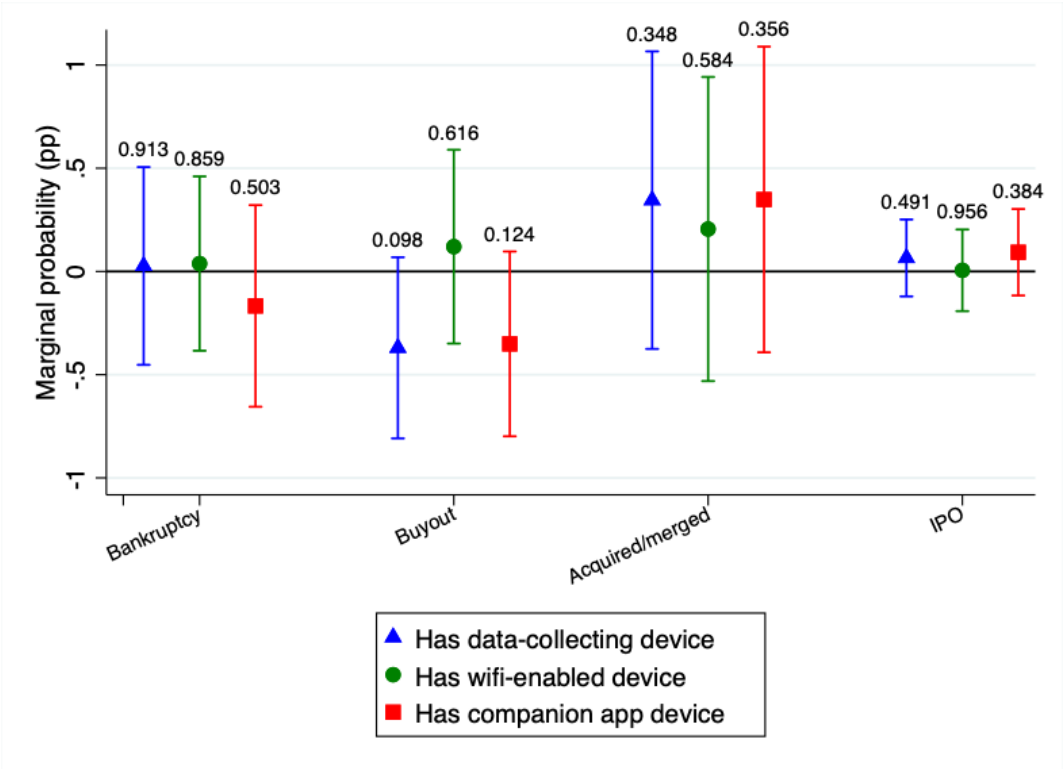
By contrast, there is no difference between early-stage data-collectors and non-data collectors after 2014. In this late period, where public data-collectors dominate non-data collectors (see Figure 3), early-stage data-collectors have no advantages over their non-data-collecting peers.

Figure A.4: Differential Exit Trajectories for Early-Stage Data-Collectors

(a) 2007-2013



(b) 2014-2019



A.5 Characteristics of Post-GDPR Exiters

Table A.2: Pre-GDPR Characteristics of Firms That Remain in versus Exit the Panel

Variable	Stayers (mean)	Exiters (mean)	p-value
Market Value ($M)	32257.8	971.2	0.000
Revenue ($M)	3943.8	200.8	0.000
R&D ($M)	321.1	9.0	0.000
Assets ($M)	22893.9	354.4	0.000
Device Count	2.98	0.36	0.059
Data Share	0.25	0.12	0.301
Any Data Device	0.37	0.27	0.517
EU Exposure	0.50	0.36	0.403

Notes: The table compares the end-of-2015 characteristics of consumer firms that remain in the panel after GDPR against those that exit. Differences in means are tested allowing for unequal variances.

A.6 Defining Consumer vs Non-Consumer-oriented Firms

Table A.3: Industries excluded from our sample to identify consumer-oriented firms

SIC Code(s)	Category	Description
Industrial & Computing Equipment		
SIC2 = 35	Machinery	Industrial and commercial machinery and computer equipment
3576	Networking	Computer communications equipment
3578	Point-of-sale	Calculating and accounting machines
3577	Peripherals	Computer peripheral equipment, n.e.c.[a]
Electronic Components & Semiconductors		
3674	Semiconductors	Semiconductors and related devices
3672	PCBs	Printed circuit boards
3679	Components	Electronic components, n.e.c.
Instruments & Medical Devices		
3812, 3823–3829	Instruments	Laboratory, scientific, and measuring instruments
3841–3844	Medical	Surgical and medical instruments and devices
Aerospace & Defense		
3721, 3724, 3728	Aircraft	Aircraft, engines, and parts
3760, 3769	Defense	Guided missiles, space vehicles, and parts
Other Manufacturing		
3714	Auto parts	Motor vehicle parts and accessories
3621	Motors	Motors and generators
3640	Lighting	Electric lighting and wiring equipment
Telecommunications & Carriers		
4812, 4813	Telecom carriers	Telephone and telegraph apparatus
3661, 3663, 3669	Telecom equipment	Communications equipment[b]
Business Services		
7371	Software	Computer programming and data processing
7380, 7389	Services	Miscellaneous business services, n.e.c.
Primary Industries, Transport, & Wholesale		
SIC2 = 10, 12–14	Mining	Metal, coal, oil, and nonmetallic mineral mining
SIC2 = 40–42, 44–47	Transport	Railroad, trucking, water, air, and pipeline transport
SIC2 = 49	Utilities	Electric, gas, and sanitary services
SIC2 = 50	Wholesale	Wholesale trade—durable goods

[a] Turtle Beach Corp (SIC 3577) is retained in the consumer sample. [b] Telecom. equipment manufacturers are excluded from both the consumer and non-consumer samples.

A.7 Additional Results

Table A.4 decomposes the financial and patenting responses by the size of treated firms' pre-GDPR data-device portfolios, splitting treated consumer firms at the treated-firm median device count of two while retaining all control firms in every column. The Tobin's Q markdown is concentrated entirely in treated firms with above-median device portfolios, with no repricing of firms holding only one or two data-capable devices. The contraction in the level of R&D concentrates among the modest-portfolio adopters, but the decline in R&D intensity appears in both groups, indicating a broad-based investment response once firm scale is held fixed. The citation-weighted AI/ML patenting decline likewise appears in both groups.

Table A.4: GDPR Effects by the Size of Firms' Pre-GDPR Data-Device Portfolios

	Treated: $\leq$ median portfolio		Treated: $>$ median portfolio	
Outcome	EU $\times$ Post $\times$ Data	Firms	EU $\times$ Post $\times$ Data	Firms
IHS Revenue	-0.464*	134	-0.080	131
	(0.246)		(0.180)	
IHS R&D	-0.844***	118	-0.218	118
	(0.236)		(0.237)	
R&D Intensity	-0.435**	110	-0.290	113
	(0.185)		(0.181)	
IHS Tobin's Q	0.341	134	-0.494**	131
	(0.349)		(0.211)	
Patents (Poisson)	-0.508	86	0.048	82
	(0.359)		(0.288)	
AI/ML Patents (Poisson)	-0.750	32	-0.379***	26
	(1.545)		(0.115)	
CW AI/ML Patents (Poisson)	-4.508***	27	-2.261***	22
	(1.725)		(0.231)	

Notes: The table re-estimates the binary triple-difference of Equation 1 on the consumer sample, keeping all control firms in every column but splitting treated firms by the size of their pre-GDPR data-device portfolios. The split is assigned once, before any outcome-specific sample restriction, at the treated-firm median device count of two: 35 treated firms fall at or below the median and 31 above, and a firm's group does not vary across outcomes. Financial outcomes are estimated on the quarterly panel with firm and quarter-by-industry fixed effects; patent outcomes are annual Poisson estimates with firm and year-by-industry fixed effects; standard errors are clustered by firm. Cluster counts differ across columns because the columns keep different treated subgroups and because firms with missing outcomes or singleton fixed-effect cells drop at estimation. R&D intensity is quarterly R&D spending divided by the firm's average quarterly revenue in 2015, with ratios above ten dropped and the remainder winsorized at the 1st and 99th percentiles. Poisson columns retain only firms with variation in the outcome. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: GDPR Exposure and Patenting per Additional Data-Capable Device

Panel A: Consumer					
	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Count	0.023***	-0.002*	0.044***	0.000	-0.000
	(0.003)	(0.001)	(0.011)	(0.002)	(0.013)
Dep. var. mean	219	27	407	1595	4181
Dep. var. SD	837	110	2450	4052	13938
Observations	949	334	933	275	949
Firms	110	45	108	37	110
Panel B: Non-Consumer					
	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Count	-0.007	0.048	0.004	0.025	0.016
	(0.012)	(0.042)	(0.015)	(0.140)	(0.013)
Dep. var. mean	143	5	98	252	2109
Dep. var. SD	403	16	285	743	9895
Observations	1,245	465	1,202	366	1,245
Firms	144	56	138	43	144

Notes: The table reports Poisson pseudo-maximum-likelihood estimates of the triple-difference coefficient, with the treatment measured as the count of pre-GDPR data-capable devices. The sample is an unbalanced panel of firms that enter by the first quarter of 2015. All regressions include firm and year-by-industry fixed effects, and standard errors are clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.8 Robustness to Balanced Panel Restriction

Table A.6: GDPR Exposure and Financial Outcomes among Firms Present Throughout 2015–2019

Panel A: Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Data	-0.341	-0.240	-0.539**	-0.365**	-0.054
	(0.267)	(0.187)	(0.216)	(0.169)	(0.232)
Dep. var. mean	8.55	6.39	3.58	0.22	1.32
Dep. var. SD	2.60	2.70	2.47	0.57	0.83
Observations	4,934	5,072	3,396	3,327	4,934
Firms	147	147	134	133	147
Adj. R-sq	0.963	0.980	0.977	0.603	0.768
Panel B: Non-Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Data	0.319	0.245	0.067	0.024	-0.056
	(0.222)	(0.165)	(0.146)	(0.028)	(0.142)
Dep. var. mean	8.41	6.49	3.79	0.13	1.18
Dep. var. SD	2.38	2.26	2.03	0.15	0.58
Observations	5,618	5,777	4,670	4,668	5,618
Firms	167	167	157	156	167
Adj. R-sq	0.968	0.980	0.968	0.847	0.750

Notes: The table reports the coefficient on the triple interaction of EU exposure, the post-GDPR period, and the firm's data-capability treatment from Equation 1, estimated on a quarterly panel of consumer device firms (Panel A) and non-consumer device firms (Panel B). The treatment is an indicator for any pre-GDPR data-capable device, and the sample is restricted to firms observed in every quarter from 2015 through 2019. All regressions include firm and quarter-by-industry fixed effects, and standard errors are clustered by firm. R&D intensity is quarterly R&D spending divided by the firm's average quarterly revenue in 2015, so that the denominator is fixed before the regulation; ratios above ten are dropped and the remainder is winsorized at the 1st and 99th percentiles within each panel. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Financial Outcomes per Additional Data-Capable Device among Firms Present Throughout 2015–2019

Panel A: Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Count	-0.028	-0.011	-0.032***	-0.018	-0.024*
	(0.020)	(0.010)	(0.011)	(0.011)	(0.013)
Dep. var. mean	8.55	6.39	3.58	0.22	1.32
Dep. var. SD	2.60	2.70	2.47	0.57	0.83
Observations	4,934	5,072	3,396	3,327	4,934
Firms	147	147	134	133	147
Adj. R-sq	0.963	0.980	0.976	0.598	0.769
Panel B: Non-Consumer					
	(1)	(2)	(3)	(4)	(5)
	IHS MktVal	IHS Rev	IHS R&D	R&D Intensity	IHS Tobin's Q
EU × Post × Count	0.042***	0.025***	0.022***	0.002**	-0.008
	(0.012)	(0.005)	(0.005)	(0.001)	(0.008)
Dep. var. mean	8.41	6.49	3.79	0.13	1.18
Dep. var. SD	2.38	2.26	2.03	0.15	0.58
Observations	5,618	5,777	4,670	4,668	5,618
Firms	167	167	157	156	167
Adj. R-sq	0.969	0.980	0.968	0.848	0.750

Notes: The table reports the coefficient on the triple interaction of EU exposure, the post-GDPR period, and the firm's data-capability treatment from Equation 1, estimated on a quarterly panel of consumer device firms (Panel A) and non-consumer device firms (Panel B). The treatment is the count of the firm's pre-GDPR data-capable devices, and the sample is restricted to firms observed in every quarter from 2015 through 2019. All regressions include firm and quarter-by-industry fixed effects, and standard errors are clustered by firm. R&D intensity is quarterly R&D spending divided by the firm's average quarterly revenue in 2015, so that the denominator is fixed before the regulation; ratios above ten are dropped and the remainder is winsorized at the 1st and 99th percentiles within each panel. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: GDPR Exposure and Patenting among Firms Present Throughout 2015–2019

Panel A: Consumer

	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Data	0.027	-0.372***	-0.217	-2.240***	-0.287
	(0.302)	(0.115)	(0.458)	(0.220)	(0.247)
Dep. var. mean	228	27	424	1595	4355
Dep. var. SD	854	110	2500	4052	14200
Observations	911	334	895	275	911
Firms	104	45	102	37	104

Panel B: Non-Consumer

	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Data	-0.325	3.568***	-0.080	3.687***	0.491**
	(0.460)	(1.118)	(0.339)	(1.194)	(0.234)
Dep. var. mean	145	5	99	255	2158
Dep. var. SD	407	16	287	746	10011
Observations	1,215	461	1,180	362	1,215
Firms	139	55	134	42	139

Notes: The table reports Poisson pseudo-maximum-likelihood estimates of the triple-difference coefficient on annual patenting outcomes, restricting the sample to firms observed in every quarter from 2015 through 2019. All regressions include firm and year-by-industry fixed effects, and standard errors are clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: Patenting per Additional Data-Capable Device among Firms Present Throughout 2015–2019

Panel A: Consumer					
	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Count	0.023***	-0.002*	0.044***	0.000	-0.000
	(0.003)	(0.001)	(0.011)	(0.002)	(0.013)
Dep. var. mean	228	27	424	1595	4355
Dep. var. SD	854	110	2500	4052	14200
Observations	911	334	895	275	911
Firms	104	45	102	37	104
Panel B: Non-Consumer					
	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Count	-0.007	0.048	0.004	0.025	0.016
	(0.012)	(0.042)	(0.015)	(0.140)	(0.013)
Dep. var. mean	145	5	99	255	2158
Dep. var. SD	407	16	287	746	10011
Observations	1,215	461	1,180	362	1,215
Firms	139	55	134	42	139

Notes: The table reports Poisson pseudo-maximum-likelihood estimates of the triple-difference coefficient, with the treatment measured as the count of pre-GDPR data-capable devices and the sample restricted to firms observed in every quarter from 2015 through 2019. All regressions include firm and year-by-industry fixed effects, and standard errors are clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.10: GDPR Exposure and Patenting after Excluding the Most Prolific Patenters

Panel A: Consumer					
	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Data	-0.630	1.145	-0.592**	-3.433***	-0.665**
	(0.402)	(.)	(0.300)	(1.232)	(0.291)
Dep. var. mean	78	6	308	313	2379
Dep. var. SD	159	17	2460	1058	6724
Observations	895	229	879	216	895
Firms	104	35	102	30	104
Panel B: Non-Consumer					
	(1) Patents	(2) AI/ML	(3) CW Pat	(4) CW AI/ML	(5) Val Pat ($)
EU × Post × Data	-0.311	2.686***	-0.510	-0.811	0.280
	(0.479)	(0.835)	(0.360)	(0.679)	(0.182)
Dep. var. mean	76	2	55	152	1090
Dep. var. SD	183	8	133	596	3509
Observations	1,200	420	1,157	293	1,200
Firms	139	51	133	38	139

Notes: The table reports Poisson pseudo-maximum-likelihood estimates of the triple-difference coefficient on annual patenting outcomes, excluding the five most prolific patenters from each panel. All regressions include firm and year-by-industry fixed effects, and standard errors are clustered by firm. Firm + Year-by-SIC2 FE. SEs clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.11: GDPR and the Composition of AI Patenting: Non-Consumer Firms

Patent category	EU × Post × Data	SE	Firms	Observations
Any AI (all components pooled)	0.329	(0.562)	116	1,012
Machine learning	2.325***	(0.407)	67	589
Evolutionary computation	0.541	(0.566)	57	507
Natural language processing	0.366	(1.084)	55	485
Speech recognition	-0.025	(0.792)	39	332
Computer vision	0.309	(0.862)	78	684
Planning and control	0.951***	(0.361)	96	837
Knowledge processing	1.674***	(0.581)	71	625
AI hardware	0.198	(0.316)	86	754

Notes: The table reports Poisson pseudo-maximum-likelihood estimates of the triple-difference coefficient from Equation 1 on annual counts of AI patents, where patents are classified into AI and its component technologies by the USPTO Artificial Intelligence Patent Dataset's text-based models rather than by CPC codes. The first row pools all AI components; the remaining rows estimate the same specification separately for each component. The sample is the unbalanced panel of non-consumer device firms that enter by the first quarter of 2015. All regressions include firm and year-by-industry fixed effects, and standard errors are clustered by firm. Poisson estimation retains only firms with variation in the outcome, so cluster counts vary across rows. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.