Data Capital: Evidence from Connected Devices*

Tarikua Erda† Ryan Gilland‡ Ben Lahey§

February 2, 2026

[Latest version]

Abstract

Data is a key form of intangible capital, yet its unobserved nature has limited our understanding of its value. Applying natural language processing to understudied regulatory documents covering the universe of wireless devices sold in the U.S., we identify and study data capability from connected "smart" devices that collect and transmit user data via apps and cloud servers. Prior commercialization of such data-capable devices is associated with lumpy but prolific subsequent product-introduction episodes and, among startups, a higher likelihood of successful exits. Data-capable public firms are 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, we find suggestive evidence that the revenues and market value of data-capable firms exposed to restrictions on data transmission under the European General Data Protection Regulation suffer, consistent with first-party data streams from connected devices commanding expected rents. Our approach offers a new, scalable measure of data capital and highlights its implications for firm value and performance in the AI era.

Keywords: Data Economy; Intangible Capital; Product Innovation; R&D; Internet of Things; Corporate Finance
JEL Codes: E22; G30; O30

*We thank Anton Babrov, Olalekan Bello, Michael Ewens, Xavier Giroud, Robert Seamans, and James Traina for helpful discussions. Support from the Alfred P Sloan Foundation is gratefully acknowledged. All 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 pointing to the quality and uniqueness of startups' proprietary data as its 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 the first-order issue of how data factors into firm performance and value, 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' data-collection capability. 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 leverage an untapped administrative dataset on wireless products and use connected devices—"smart" products that can connect to apps or the cloud and continuously generate, process, or transmit user data—to develop a novel measure of firms' data capability. We document new facts characterizing data-capable firms and product innovation. We also exploit a data collection cost shock created by the European Union's General Data Privacy Regulation ("GDPR"), which restricted key forms of data transmission and processing in 2018 for firms with customers in the European Economic Area. EU-exposed data-capable firms see declines in revenues and market value after this regulatory change, suggesting material expected rents linked to intangible data capability.

Connected devices make a useful setting for two reasons. First, the booming growth in these devices defines the modern, Internet of Things ("IoT") era of 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 only projected to grow in the rest of this decade.[2] Second,

[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, are only available for sampled firms/years.

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

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 be reused to improve products, develop complementary products, and encourage ecosystem lock-in. 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.

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—often routed through a companion app and cloud service. The firm can use these data to improve navigation and reliability, inform product iteration, support add-on services such as software features, replacement parts, and more. 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 on 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, which is the last year of data where we have detailed and comprehensive data on the FCC website.

We document that data-capability is associated with lumpy but prolific patterns in

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

[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.

product innovation. While new-device filing episodes are generally rare, FCC grantees that have previously commercialized data-capable devices subsequently introduce greater numbers of new devices when they do file, consistent with platform-style development where firms batch releases and iterate around a data-enabled core.

We then use Compustat and Pitchbook to characterize public and private 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, including in sector-adjusted terms. They have dominated patenting activity in AI and machine learning technology areas, where data is a critical input. Among early-stage private firms, bringing a data-capable device to market was associated with much higher likelihoods for successful exits, particularly IPOs. However, this advantage that waned starting in the mid-2010s, concurrent with the broader diffusion of IoT technologies.

Finally, we use a triple difference design to quantify 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 the GDPR regulation is put into force in 2016 relative to non-data capable firms with EU exposure. This approach is helpful for identification because, given that firms with data capability are systematically different, it helps to isolate the GDPR's impact from confounders like broader differences in firm technology, managerial quality, and growth opportunities.

We show GDPR is associated with economically meaningful declines in firm value among firms that were *ex ante* data capable. In our preferred specifications with firm andindustry by year fixed effects, firms that, before GDPR, were EU-exposed and data-capable experience larger declines in market value (by more than 40%) and revenue (by about a quarter). We also find suggestive evidence of declines in R&D spending and Tobin's Q following the GDPR. These findings suggest that data is valuable and markets reprice when shocks erode its usability. In future work, we plan to expand the coverage and matching of public and private firm data to the FCC grantee database to better understand these impacts.

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 the context of first-party data collection

via connected devices, and document declines in not only revenues, but also market values after GDPR, reflecting the value of data as an intangible asset.

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), or 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, methodologically, we also add 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 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

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

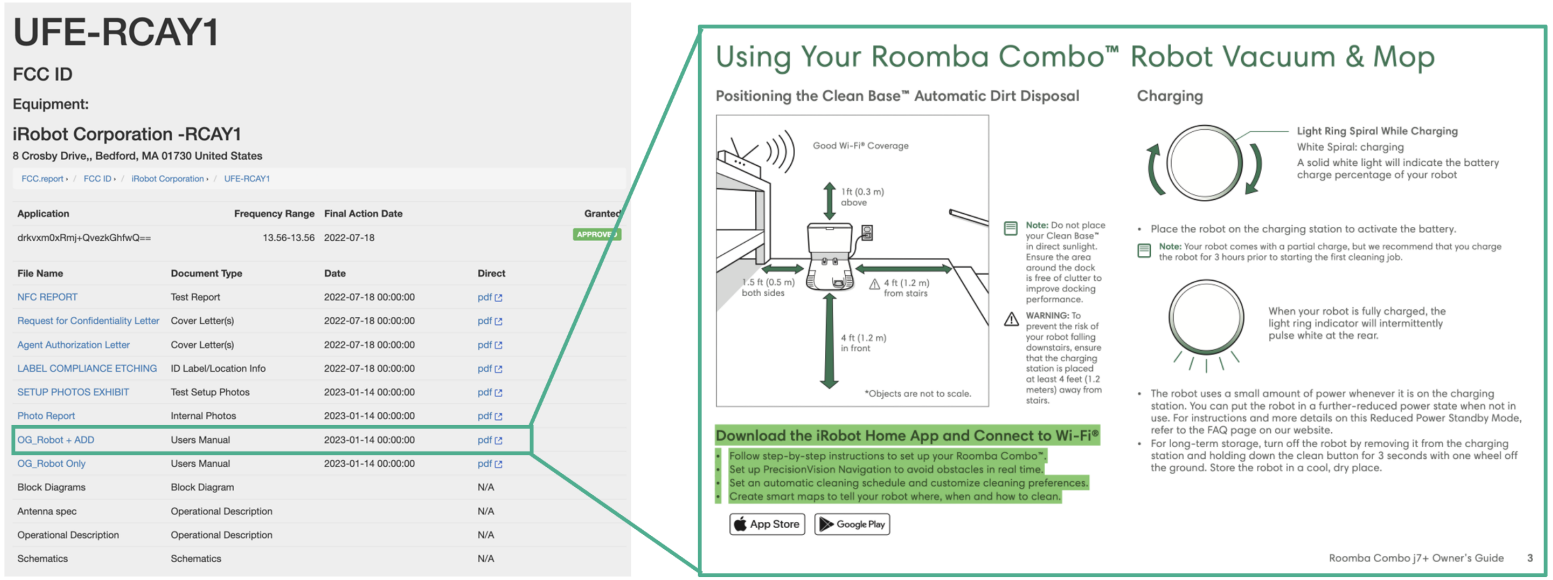
of process, ensuring each authorized device is linked to a U.S.-jurisdictional entity. Because authorization precedes market access, these filings are mandatory and independent of firms' financial reporting considerations, performance, or strategic disclosure choices.

We build a device-level FCC dataset by scraping the metadata and application documentation on user manuals associated with device filings.[5] 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 economically meaningful device features—such as whether a product supports persistent connectivity, data transmission, or companion mobile applications—that distinguish simple RF devices (e.g., garage door openers) from data-capable connected products (e.g., smart wearables and home appliances).

As an illustrative example, Figure 1 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.

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

Figure 1: 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

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.

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.

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

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, research and development (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 requirements 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 PricewaterhouseCooper 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 [6].

In examining the effects of connected device introduction on longer-run firm outcomes, we leverage deal data from PitchBook. We consider a variety of financing outcomes, in particular initial public offerings, strategic acquisitions, and private equity-led buyouts. While the PitchBook universe extends across public and private firms, complimentary financial data for private firms is extremely limited.

[6] The perspectives of Deloitte and PricewaterhouseCooper can be viewed through these links.

Patent Information

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. In subsequent analyses we emphasize the date a patent was first filed with the USPTO, rather than the date it was granted, to account for a notable lag in the time between filing and approval.

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.

2.1 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 Compustat and PitchBook databases. Details of the matching algorithm are described in the appendix. Overall we match 3,286 (11.9 percent) of the uniquely identified firms in the FCC device filings to those in Compustat, and 9,791 (35.3 percent) to those in PitchBook. The disparity in match rates between Compustat and PitchBook is unsurprising, as any firm can file with the FCC in principle. The firms captured in Compustat are current or former publicly-listed entities, while PitchBook's scope is far broader and aims to capture the universe of "investable" firms.

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 1,139,053 (33.3 percent) of the patents in our initial sample, of which 9,820 are classified as AI/ML. We then build quarterly 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 any sales exposure 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 150 of the FCC-Compustat linked firms as being potentially exposed to GDPR. An additional 319 firms report segment information but do not identify an EU-based geography.

2.2 Validating our Measure of Data Capability

To validate our measure of firms' data-collecting capability, we leverage a separate, external resource—the texts of purchase price allocations in Securities Exchange Commission (SEC) filings. In particular, we examine whether acquisitions 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 178 deals, where either the target or acquiring firm sold FCC-licensed devices. Of these, 29 deals specifically linked to target firms.[7] 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 reas of user engagement and retention metrics, IoT and connected devices (e.g., wearables like Fitbits). Lastly, we test for statistical significance using Fisher's exact test for a binary indicatory of *any* keyword presence.

[7]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.

Our findings in Table 1 reveal that acquisitions involving data-device companies exhibit significantly higher mentions of data-related strategy keywords across multiple categories. Strikingly, a binary indicator for *any* keyword presence shows that 54.5% of data-capable targets have at least one data-strategy keyword mentioned, compared to only 11.1% of non-data capable targets (odds ratio = 9.6, p=0.028). Looking at all deals whether either the target or acquirer was a data-capable company similarly shows a large increase in the likelihood of data-use or data-strategy mentions in the deal narrative—mentions that, according to the detailed breakdown of keyword types in Appendix Table A.1—appear to be primarily associated with engagement/retention terms, IoT terminology, and data collection language. When restricting to the smaller set of deals where the *target* (rather than acquirer) is the data-device company, the IoT category still remains significant. These findings suggest that FCC data-device classifications capture meaningful variation in how acquirers discuss data assets and strategy in their SEC disclosures, and thus lend credence to our measurement of the data capability of firms is reasonable.

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

	Target-Only (N=29)			**All Deals (N=178)**		
	Has Keyword	Total	%	Has Keyword	Total	%
Data-device deals	6	11	54.5%	49	121	40.5%
No-data-device deals	2	18	11.1%	11	57	19.3%
Difference			+43.4pp			+21.2pp
Odds Ratio		9.60			2.85	
p-value (Fisher's exact)		0.0281**			0.0063***	

Notes: 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).
*** p<0.01, ** p<0.05, * p<0.10

3 Characterizing Data-Capable Devices and Firms

3.1 Data-Capable Devices and Product Innovation

Table 2 shows how data capability factors into product innovation, as inferred from new product filings. In particular, we estimate the effect of prior data capability devices on new product filings by constructing a balanced panel for the 23,212 FCC grantees observed from 2007 to 2019. Column 1 shows that firms with prior data device experience are 4.8 percentage

points less likely to file any new product, relative to a baseline probability of 30.6%. In terms of total counts of new device filings, however, we find an increase of 0.102 new device filings among data-capable firms, representing a 13.7% increase relative to the baseline mean of 0.75 new products per firm-year. That a lower filing probability coexists with a higher unconditional device count implies that data-capable firms' filing episodes, when they occur, are more prolific, consistent with platform-style development where releases are batched around a data-enabled core.

Table 2: Data-Capability and New Product Innovation Dynamics

	Dependent Variable	
	(1)	(2)
	Any New Device	# of New Device Filings
Prior Data Device Experience	-0.048^{***}	0.102^{***}
	(0.006)	(0.035)
Firm Age	-0.229^{***}	-0.317^{***}
	(0.002)	(0.011)
Firm Age^2	0.015^{***}	0.014^{***}
	(0.000)	(0.001)
IHS(Prior Products)	-0.186^{***}	-0.614^{***}
	(0.005)	(0.044)
Baseline Mean	0.306	0.746
Observations	140,510	140,510

This table reports OLS estimates of the effect of prior data device experience on product innovation. The sample is a balanced panel of FCC grantees from 2007–2019. Column (1) estimates an LPM for the probability of introducing any new product. Column (2) estimates the regression for the total number of new products filed as an outcome. *Prior Data Device Experience* is a binary-indicator equal to one if the firm filed at least one data-capable device in any prior year. The control for firm age (linear and quadratic) is calculated as years since the grantee's first FCC filing, and helps account for lifecycle dynamics in product filings. The inverse hyperbolic sine of cumulative prior products controls for the grantee's filing history while accommodating zeros. All specifications include grantee fixed effects and interacted year fixed effects with country, equipment class, and product function. Standard errors clustered by grantee are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.2 Data-Capability among Public Firms

Figure 2 shows various summary statistics comparing public firms with at least one data-capable device against 'non-capable' firms with none. Panel (a) clearly shows that, the share of data-capable firms has risen sharply: the number of data-capable public firms overtook

non-capable firms by the mid-2010s. This is consistent with a broad shift toward connected products since the introduction of smartphones and apps. Data-capable firms are larger and more highly valued: their median assets and market value exceed those of non-capable firms, with the gap widening after 2014 (see Panels b and c). Median Tobin's Q tracks closely across the two groups for much of the period (Panel d).

Notably, data capable firms also appear more innovative. According to Panels (e) and (f), data capable firms are also more R&D intensive, both in absolute spending and relative to revenue, consistent with data from connected devices serving as an input into product improvement and development. Perhaps most strikingly, data-capable firms have come to dominate patenting activity—particularly in AI and machine learning, where the gap is enormous and widens dramatically after 2015 (Panels g, h). This is consistent with data capability and AI innovation being complements: firms that generate first-party data through their devices are disproportionately the ones investing in the prediction technologies that use it.

Taken together, these findings suggest that data capability has been increasingly characteristic of larger, more R&D-intensive, and more innovative public firms. These patterns generally also hold in sector-adjusted comparisons, but divergence weakens over time as technologies diffuse. This suggests that we are capturing not just across-sector differences in size and performance, but real differences among data capable and non-capable firms in similar industries.

3.3 Data-Capability among Early-Stage Firms

We estimate a discrete hazard model from Equation 1 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

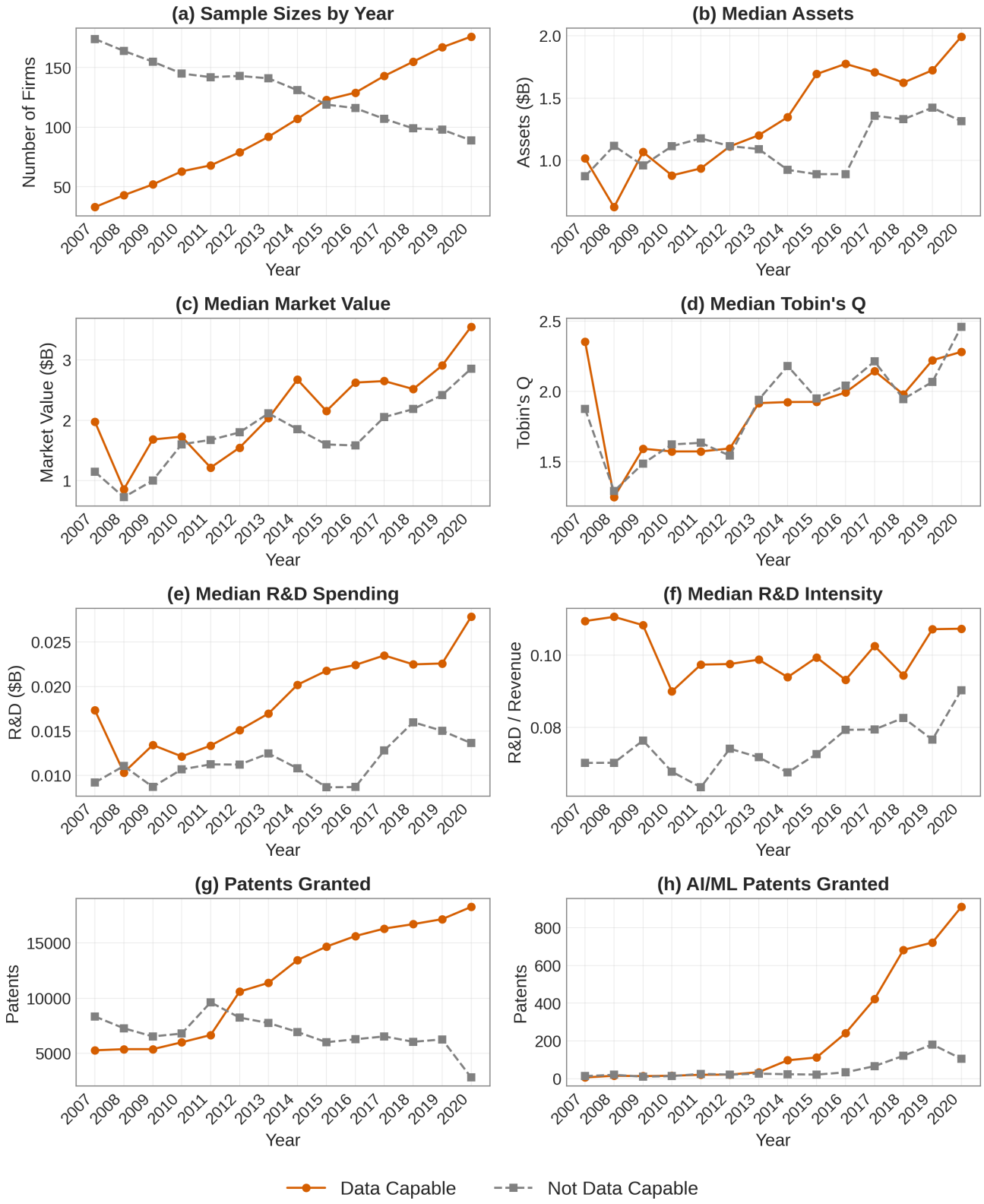
{first PE deal, first corporate deal, acquisition, bankruptcy, IPO}

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 number of devices of all types. where $X \in$ {data-collecting, has companion app}.

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

The control term C contains firm age bins, deal history (indicators for whether the

Figure 2: Comparing data capable versus non-capable public firms (2007-2020)



firm has ever secured venture, PE, and corporate deals), and indicators for most recent firm status observed as of t: profitability, revenue generation, trials/development, and start-up. 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.

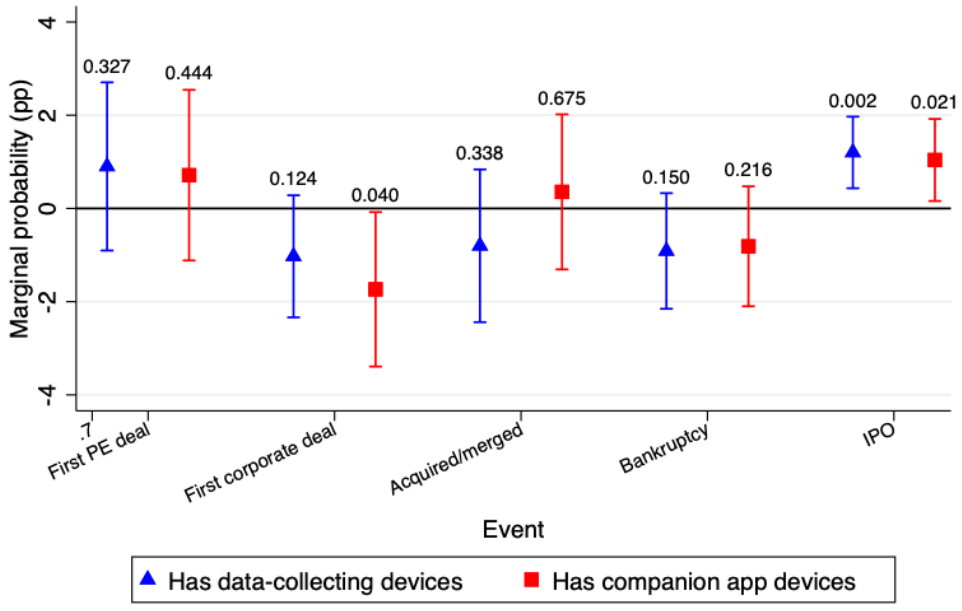
Figure 3 plots β_1 for each outcome from Equation 1 from 2007-2013 in Panel A and from 2014-2019 in Panel B. We split the sample in 2014 to consider the trajectories of data-collecting firms in two periods that appear distinct; as Figure 2 demonstrates, 2014 is the first year in which the median data-collector had more assets, a higher valuation, and spent more on R&D than the median non-data-collector.

Figure 3 shows that – conditional on age, deal histories, profitability, and fixed effects – both firms with data-collecting devices and firms with companion app devices were about 1 percentage point more likely to IPO than firms without, and firms with companion app devices were nearly 2 percentage points less likely to secure their first corporate deal. In the cross-section without controls, 8.9% of data-collecting firms before 2014 ever IPO (including in years after 2014) while just 0.8% of non-collectors ever IPO. This difference likely reflects the fact that data-collecting and companion app firms were relatively cutting-edge before 2014.

By contrast, there is little 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 2), early-stage data-collectors have no advantages over their non-data-collecting peers aside from a 1.5 percentage point higher probability of securing a first PE deal (pvalue < 0.1). In the cross-section without controls, 7.2% of collectors ever IPO while 3.6% of non-collectors ever IPO, but this difference is erased upon the addition of deal history and profitability controls.

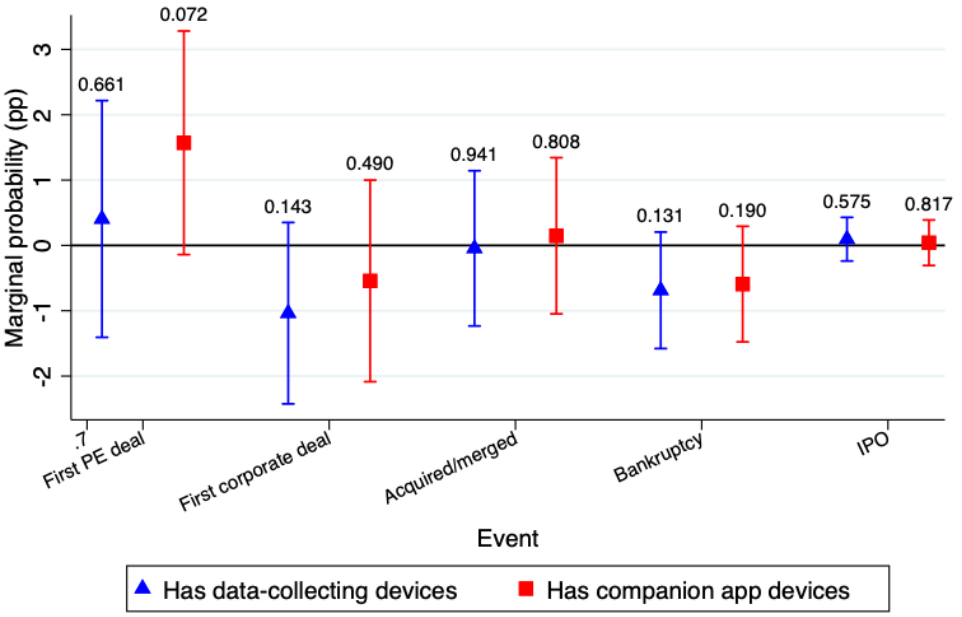
Figure 3: Differential Financing & Exit Trajectories for Early-Stage Data-Collectors

(a) 2007-2013



Controls for log filed devices, age bins, previous deal indicators, profitability/revenue/development statuses, and state and sectorXyear FEs. SEs firm-clustered.
Considers only obs up to first event among US firms with devices.

(b) 2014-2019



Controls for log filed devices, age bins, previous deal indicators, profitability/revenue/development statuses, and state and sectorXyear FEs. SEs firm-clustered.
Considers only obs up to first event among US firms with devices.

4 Data Capability and Firm Outcomes

4.1 Background on GDPR

The EU's General Data Protection Regulation (GDPR)—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 core processing principles—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 Euros million 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.[8]

4.2 Estimation and Results

We implement a triple differences-in-differences design from Equation 4.2 on a balanced 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 2023q4 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 changes in anticipation of the GDPR during the phase-in period in the overall ATT of the GDPR. We restrict the sample to firms active in all quarters from 2015q4-2023q4.

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

$$
\begin{aligned}
Y_{it} = \beta_1[\mathbb{1}(t \geq t^*) \times \mathbb{1}_i(\text{EU-exposed}) \times \mathbb{1}_i(\text{has data collecting devices})] \\
+ \beta_2[\mathbb{1}(t \geq t^*) \times \mathbb{1}_i(\text{EU-exposed}] \\
+ \beta_3[\mathbb{1}(t \geq t^*) \times \mathbb{1}_i(\text{has data collecting devices})] \\
+ \beta_4[\mathbb{1}_i(\text{EU-exposed}) \times \mathbb{1}_i(\text{has data collecting devices})] \\
+ \lambda_i + \eta_{year \times s(i)} + \epsilon_{it} \quad (2)
\end{aligned}
$$

We consider as outcomes Y_{it} the firms' quarterly log market value, log revenue, log R&D spending, and Tobin's Q. Equation 4.2 includes firm fixed effects λ_i and year $\$times$ 2-digit SIC industry $\eta_{year \times s(i)}$ 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 × 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 flat pre-trends in our event study plots in Figure 4, suggesting a lack of differential evolution between groups leading into GDPR. The primary threat to (A2) is that the GDPR drove differential compositional changes across groups, so in our primary specifications, we restrict our sample to the set of firms active throughout 2015q4-2023q4 to ensure our results are not driven by entries and exits. (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 year × sector fixed effects. The EU-exposure measure is based on customer segment information described in 2.1 and we validate our data-collection measure in 2.2.

Table 4.2 presents β_1, β_2, and β_3 from Equation 4.2. The triple difference coefficients in Column (1) and Column (2) are significant and suggest that the GDPR was devastating for EU-exposed data-collectors. These estimates correspond to a 43% decline in market

value and a 27% decline in revenue over the pooled post-GDPR period. While estimated imprecisely, the negative coefficients on log R&D spending and Tobin's Q suggest that .

Table 3: Triple Difference Estimates: GDPR Effects on Data-Capable Firms

	(1) Log Market Value	(2) Log Revenue	(3) Log R&D	(4) Tobin's Q
EU Exposure × Post	0.143	-0.003	-0.075	0.101
	(0.161)	(0.124)	(0.113)	(0.325)
Post × Data Device	0.193	0.113	-0.057	0.066
	(0.130)	(0.095)	(0.109)	(0.426)
EU × Post × Data Device	-0.560**	-0.308*	-0.169	-0.544
	(0.263)	(0.167)	(0.184)	(0.768)
Observations	13,071	14,817	9,909	13,068

Standard errors clustered at firm level in parentheses.

All regressions include firm and year x industry fixed effects.

EU x Data Device absorbed by firm FE (time-invariant).

Post period: 2016–2024 (following announcement of GDPR in 2016q2).

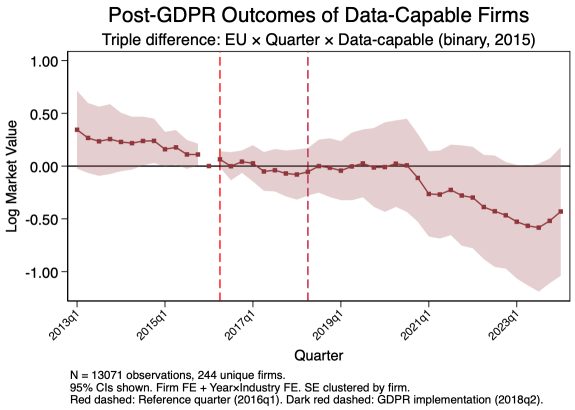
Data device indicator measured in 2015 (or most recent pre-2015 year).

We estimate the evolution of β_1 about the GDPR using the framework from Equation 4, but interacting the triple-difference terms with indicators for each quarter to form event studies. We take 2016q1, the quarter before GDPR passed the European Parliament, as the reference quarter and set this coefficient to 0.

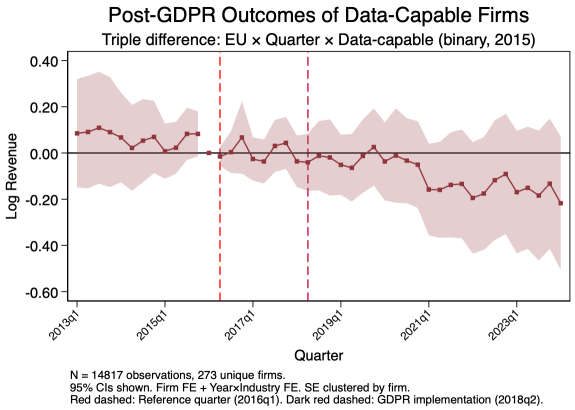
Figure 4 suggests that the financial effects of the GDPR were not felt until some time after the regulation was implemented, which is consistent with an explanation of these effects as declines in data-reliant firms' abilities to target, refine, and innovate their product lines using consumer data. The pre-GDPR trends in market value do appear to drift down, though (A1) is not obviously violated since the downward drift is followed by a long, flat period before the post-GDPR value decline. In the market value and revenue plots, the first signs of weakening appear in 2020q4 and 2021q1; together with our inclusion of year × industry fixed effects this timing alleviates concerns that our effects are driven by the COVID-19 pandemic, which took profound effects on firm values in early 2020 that do not appear in our event studies. However, we cannot rule out an added, compounding negative impact of the COVID-19 downturn in later periods. Panel C shows of a decline in R&D spending for GDPR-exposed firms, suggesting a shrinking of their research and development activity following the GDPR shock that created new barriers to consumer-data, which is

Figure 4: Triple-Difference Event Studies Around the GDPR

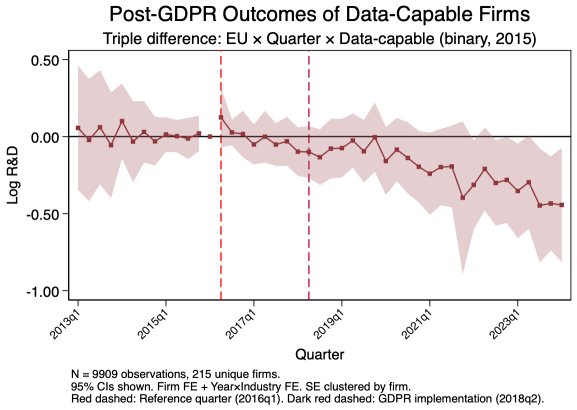
(a) Log Market Value



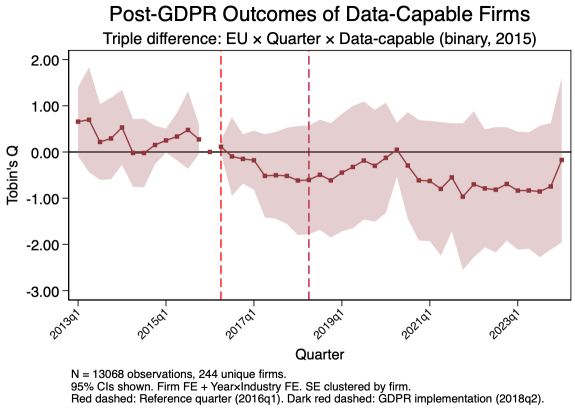
(b) Log Revenue



(c) Log R&D Spending



(d) Tobin's Q



plausibly an input into data-capable firms' R&D projects.

5 Discussion and Conclusion

This paper studies the measurement and value of data capability as an intangible asset in the Internet-of-Things era since the late 2000s. We link firms to the wireless products they commercialize using FCC equipment-authorization filings and apply NLP to identify data-capable "smart" devices—products that can collect, transmit, and analyze user data via companion apps and cloud services. We document novel stylized facts on how firms which sell data-capable devices subsequently exhibit few but prolific new device introduction episodes. Data-capability, as well as patenting activity in data-intensive domains like AI and machine learning, have been primarily concentrated among larger, higher-valued, R&D intensive public firms. We also exploit the EU's General Data Protection Regulation as a negative shock to the expected returns to data transmission in a triple-difference design, showing that data-capable public firms with greater European exposure experience sizable declines in market value and revenues after GDPR. This suggests that first-party data streams are highly valued by markets. Together, the results provide new insights into firms' investments in data capability through the devices they sell, and how that capability factors into firm value and performance. In future work, we plan to expand the coverage and matching of public and private firm data to the FCC grantee database to get broader insights into these dynamics.

One caveat is that we do not directly observe the data features of each device—i.e., whether and how exactly it collects, processes, and transmits data—but we generally assume that one or more of these activities are taking place based on mentions of companion apps or server connections in user manuals. While important, this does not undermine the core interpretation: what matters for valuation is often the option value created by the capability to collect data and the organizational investments that accompany that capability (e.g., building the data pipeline, software, and complementary services). FCC filings capture the moment a firm crosses a meaningful technological threshold—marketing a device with data-capable features—which can shift expectations about future product ecosystems, recurring revenues, and learning-by-doing. In that sense, the paper is designed to measure the emergence of a data-capable product strategy, not to replace demand-side diffusion metrics, which are a complementary and a promising next step.

It is worth noting here that our estimated magnitudes about GDPR impacts are large, with the revenue drop estimates are more than 4 times larger than those estimated in pre-

vious work on the impacts of GDPR (Quan, 2024). This is potentially explained by the fact that our measure of data-collection capability selects firms whose business models are centered around exactly the type of first-party data transmission that GDPR restricted. We also cannot rule out compounding negative impacts from the COVID-19 pandemic. Moreover, the magnitudes also depend on what untreated firms experience. If GDPR induced broader adoption of stricter privacy practices globally—as suggested by the wave of post-GDPR privacy legislation observed in other jurisdictions in recent years—then firms we classify as untreated would also bear negative effects, and our estimates would understate the true cost of data restrictions. On the other hand, we cannot rule out that competitive reallocation benefited non-treated firms at the expense of EU-exposed data collectors, which would work in the opposite direction. The net sign of these spillovers is ultimately an empirical question that our design cannot resolve, but either channel is consistent with data capability generating substantial rents that privacy regulation disrupts.

The paper is intentionally agnostic about consumer welfare and distributional effects. Data-capable devices could generate consumer surplus through better products and services, but they could also raise privacy and market power concerns such that increases in consumer welfare may offset the losses that data-capable firms face. These are beyond the scope of our study. Our contribution is to provide a clean, scalable measure of firms' data-capability investments and to document how they factor into firm value, which future work can then build on.

Our study offers a new way to measure data-related intangible capital using administrative records that are broad, standardized, and unaffected by firm strategic disclosure behavior. We document that the shift into data-capable devices is associated with meaningful changes in firm outcomes consistent with data as a productive asset. This work opens up several research avenues at the nexus of the data economy and intangible capital, and also informs current policy discourse around privacy regulation.

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

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

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

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

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/

A Appendix

Table A.1: Mean Keyword Frequency by Category: Target-Only vs. All Deals

Category	Target-Only (n=29)				All Deals (n=178)			
	Data (n=11)	NoData (n=18)	Diff	p-value	Data (n=121)	NoData (n=57)	Diff	p-value
Engagement / Retention	1.45	0.00	+1.45	0.248	0.42	0.04	+0.39	0.005***
IoT / Connected Devices	3.36	0.06	+3.31	0.088*	0.78	0.25	+0.53	0.044**
Data Collection	0.91	0.00	+0.91	0.235	0.25	0.00	+0.25	0.024**
Monetization	0.55	0.00	+0.55	0.341	0.09	0.00	+0.09	0.101
Advertising	2.27	0.06	+2.22	0.220	0.53	0.44	+0.09	0.815
Unit Economics	7.09	0.00	+7.09	0.341	0.75	0.09	+0.66	0.310
Personalization	0.00	0.06	−0.06	0.331	0.07	0.12	−0.05	0.546

Notes: Mean keyword mentions per deal. We scan acquisition-related disclosure text for grouped terms capturing (i) engagement and retention (e.g., engagement, retention, stickiness, DAU/MAU), (ii) connected-device and IoT functionality (e.g., sensor, telematics, wearable, smart home, remote monitoring), (iii) data collection and telemetry (e.g., usage/behavioral/location data, device/sensor data, geolocation, biometric/health data), (iv) cloud connectivity and companion apps (e.g., cloud, backend, platform, API, sync/OTA, streaming, mobile/companion app), (v) monetization and advertising (e.g., monetize, recurring revenue, ad targeting, first-party data), (vi) unit economics (e.g., ARPU, LTV, CAC, churn), and (vii) personalization and analytics/ML (e.g., personalization, recommendation engine, graphs/identity, analytics, machine learning/AI, predictive algorithms), including subscription language (e.g., subscription, SaaS). We exclude common false-positive phrases such as "data center" and "selected financial data."
*** p<0.01, ** p<0.05, * p<0.10

Figure A.1: Sector-adjusted comparisons across data capable versus non-capable public firms (2007-2020)

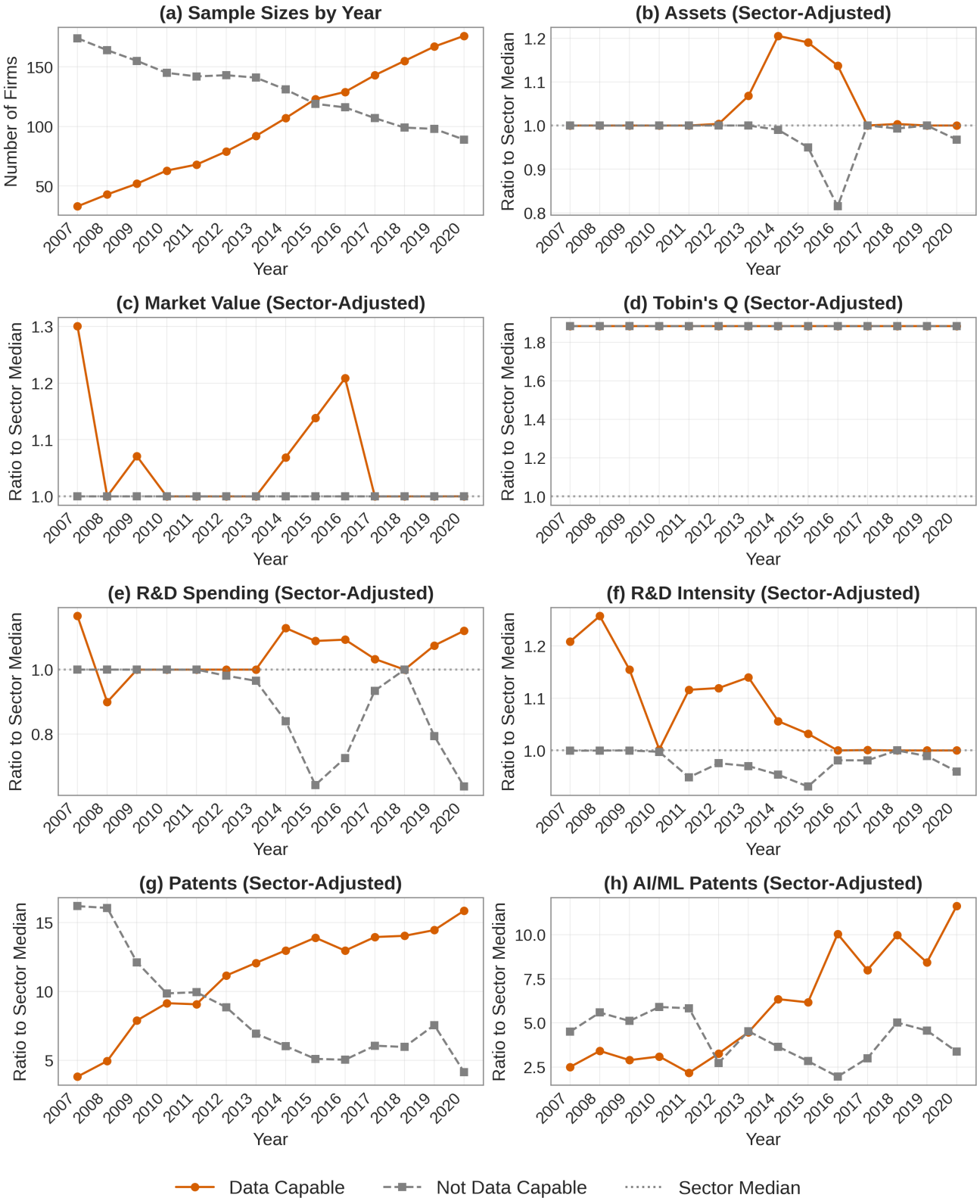
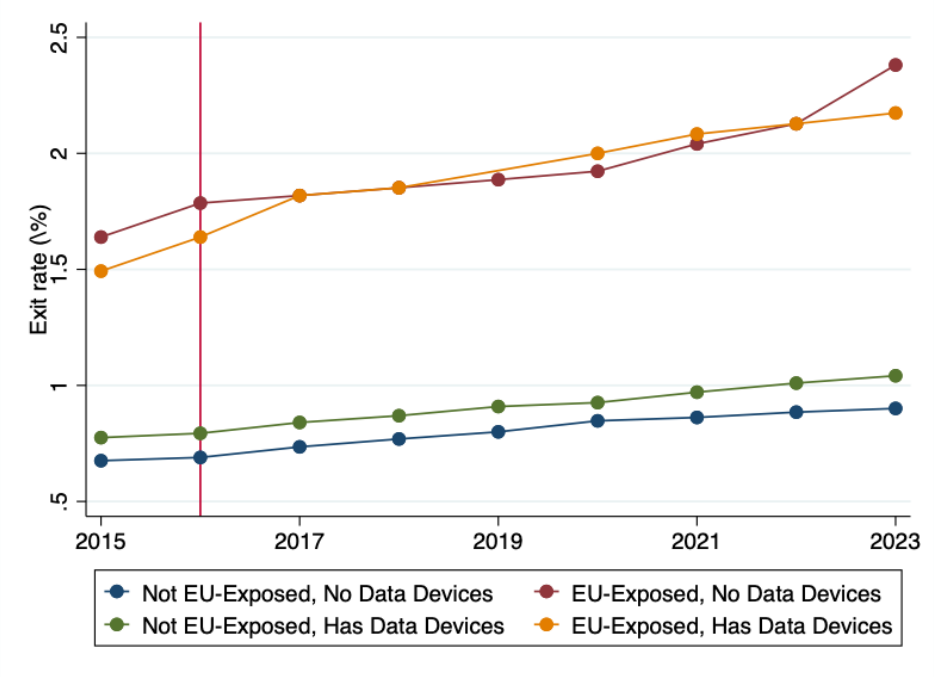


Figure A.2: Exit Rates Across Triple-Difference Design Groups



A.1 Details of the Matching Algorithm

1. For each firm-quarter in Compustat, identify the name, address, and contact details corresponding to that point in time. For PitchBook, use all provided current and past names and addresses.

2. Standardize firm details across the FCC, Compustat, and PitchBook datasets.

3. Exact match FCC grantees to Compustat or PitchBook firms on web URL and fax number.

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 all matches on web URL or fax number, 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.