Data Privacy Regulation, Compliance Costs and Startup Growth Rates*

First draft: 12 January 2026
This draft: 25 March 2026

Abstract

How do data privacy regulations affect high-growth entrepreneurship? I study the impact of the EU's General Data Protection Regulation (GDPR) on high-growth data-intensive (henceforth HGDI) startups using employment histories covering nearly ten million job spells. GDPR increased compliance hiring at EU startups relative to US counterparts: for data-intensive industries, the estimated effect accounts for roughly 10–20% of the EU–US gap in compliance staffing. GDPR is associated with slower firm growth and fewer new HGDI startups in Europe, with effects that are larger in more data-intensive industries. These findings suggest that privacy regulation can impose meaningful costs on high-growth entrepreneurs.

JEL classification: L26, L51, L53.

Keywords: GDPR, Regulation, Workforce data.

*

1 Introduction

Economic growth in Europe has lagged behind the United States in recent years, and the gap is especially visible among high-growth technology firms. While the US has produced a large number of startups that have scaled rapidly, Europe has produced fewer comparable success stories, particularly in data-intensive sectors such as AI and social networking. A prominent explanation is that regulation raises fixed costs that are disproportionately burdensome for young firms (Garicano, Holmström, and Petit, 2025). The General Data Protection Regulation (GDPR) is often singled out: by tightening rules on the collection and use of personal data, it may force startups to divert resources toward compliance rather than product development. These concerns feature in policy discussions, including the "Draghi report" on European competitiveness (Draghi, 2024), yet there is limited direct evidence on the magnitude of the compliance burden for startups. Indeed, early impact assessments by the European Commission argued that an EU-wide regime could *reduce* costs by harmonizing national privacy rules (European Commission, 2012).

This paper provides direct evidence on how GDPR changed startups' compliance investments and growth. Using employment histories from Revelio Labs covering nearly ten million job spells at startups, I track the evolution of workforce composition as firms scale and compare high-growth data-intensive startups in the EU and the US before and after GDPR's implementation. I show that GDPR increased compliance hiring at EU startups relative to US counterparts, accounting for roughly 10–20% of the post-2018 EU–US gap in compliance staffing in the most data-intensive industries. Because many compliance-related tasks are performed by non-specialists or outsourced, these estimates should be interpreted as a lower bound on the overall compliance burden.

I also study how GDPR affected startup activity and firm growth. I show that after 2016 and 2018 (GDPR's announcement and implementation years), Europe had fewer high-growth data-intensive (HGDI) startups, especially those reaching more than 100 hires within four years of founding. As this result could be consistent with numerous other explanations

such as the US being dominant in technologies such as AI due to other factors or attracting more venture capital, I show that this gap is more pronounced the more data-intensive a startup is. I also show that after the introduction of GDPR, European HGDI startups are less likely to reach key milestones, such as 50 employment spells. These findings suggest that data privacy regulations impose meaningful constraints on startup scaling by imposing an early fixed cost that reduces resources available for core staff.

My identification compares European startups founded after the announcement and implementation of GDPR to those founded before and to US startups.[1] This comparison allows me to control for differences between EU and US startups caused by other regulations - EU HGDI startups hire more compliance employees than US startups even prior to the implementation of GDPR.

There are several threats to the validity of this research design. One concern may be that the results are not driven by GDPR but instead by other European regulations. To study the effects of GDPR, which predominantly affects data-intensive startups, I use a triple-interaction model to show that the effect of GDPR increases with the data intensity of a firm. Though I find that *non-data-intensive* EU startups also increased compliance hiring after GDPR, a visual examination reveals that this effect appears to be driven by a sharp decline that started prior to GDPR and leveled off around the time of its implementation. Another concern might be that GDPR also affected American startups, for instance because of the "Brussels effect", where European regulations get adopted globally due to the importance of the EU market (Bradford, 2020). I show that this appears not to be the case - compliance shares rise more in EU firms than American ones.

The data come from Revelio Labs, which aggregates and structures publicly available LinkedIn profiles.[2] These data are particularly well-suited for studying workforce dynamics

[1]This is similar to a differences-in-differences strategy, except that individual startups do not become "treated", rather, I compare startups founded before and after GDPR's implementation. I also report very similar results when looking at job spells instead of startups.

[2]The data do not cover workers without LinkedIn profiles. I discuss this limitation in Section 3. In general, startups and knowledge workers are well-represented in the Revelio Data. Core, Ryduchowska, and Wang (2024) find matches for about 75% of startups that received venture capital financing (according to

in high-growth startups. Critically, the data include the start date of each job position, allowing me to track how workforce composition evolved as startups grew. I focus on data-intensive industries where GDPR compliance requirements are most relevant, identified using AI-based classification of 400 industry categories.

To measure the impact of GDPR on compliance costs, I study job positions. I classify job positions into compliance-related roles using AI. I start with Revelio's classification of positions into 17000 "role categories." This classification is based on job titles, work history, and educational backgrounds. I then classify these roles into compliance (for example: legal, privacy, data protection officers), core operations (engineering, sales, product), or other support functions (finance, HR, administration) using large language models.[3] As mentioned above, this approach likely understate GDPR's true impact, as "core" workers increasingly perform compliance-related task: for example, CEOs leading GDPR initiatives or engineers redesigning databases to limit personal data storage. However, comparing the EU to the US before and after GDPR allows me to see the *relative* change in compliance burdens despite the difficulty of quantifying the exact cost of compliance for any specific firm.

Prior to the implementation of GDPR, HGDI startups in Europe had about two employment spells in compliance per 100 employment spells whereas US startups had 1.5. After the implementation of GDPR, the share of compliance spells among total spells remained relatively constant among US startups. However, in Europe, the share rose from 1.9 in 2016 to 2.2 in 2022. My coefficient estimates imply that roughly 10-20% of the "compliance gap" can be explained by GDPR. The compliance gap did not widen among non-data-intensive startups, providing suggestive evidence that GDPR was the reason.

The effects on startup rates and growth rates are also economically meaningful. Prior to GDPR, HGDIs founded in Europe and the US were equally likely to reach 50 cumulative

Pitchbook and VentureXpert) between 2010 and 2017 in the Revelio data while Chen, Huang, and Li (2026) match 82% of US VC-backed startups (from Pitchbook) to Revelio. Both of these are likely to understate the true rate of VC-backed startups in the Revelio data due to problems with matching, such as inconsistent LinkedIn URLs due to name changes across the two databases.

[3]In my primary analysis, I use a classification from Anthropic's Claude (Claude 4.5 Sonnet) model.

employment spells within four years of founding (about 15% of startups reach this threshold). After the implementation of GDPR, US firms were about 2-3 percentage points more likely to reach 50 employment spells than European ones. This gap persists for higher thresholds, such as 150 employment spells, 200 employment spells and 300 employment spells. Crucially, the gap does not appear to have widened for non-data-intensive startups.

This paper contributes to several literatures. First, it adds to the growing evidence on the cost of regulations for firms (for a summary of this literature, see Bombardini, Trebbi, and Zhang (2024)). In particular, I contribute to the literature on the costs of GDPR. Johnson (2022) provides an overview of the literature on the impacts, both positive and negative, of GDPR.

Prior research has documented that GDPR implementation had negative effects on European startups. Post-GDPR, European startups received less trans-Atlantic venture capital investment (Jia, Jin, Leccese, and Wagman, 2025), stored less data (Demirer, Jiménez Hernández, Li, and Peng, 2024), launched fewer new apps (Janßen, Kesler, Kummer, and Waldfogel, 2022), and received fewer pageviews (Goldberg, Johnson, and Shriver, 2024). GDPR implementation was also associated with decreased activity among smaller companies providing website analytics tools, benefiting larger players like Google (Johnson, Shriver, and Goldberg, 2023; Peukert, Bechtold, Batikas, and Kretschmer, 2022).[4] Deisenroth, Manjeer, Sohail, Tadelis, and Wernerfelt (2024) use Apple's limitations on app tracking to draw an analogy to GDPR and show that privacy regulation is likely to cause more market concentration and hurt startup investments in marketing. These papers generally document that GDPR hurt firms by reducing profits. In contrast to these papers, I show that GDPR also hurts firms by increasing the cost of doing business. Koski and Valmari (2020) and Frey and Presidente (2024) attempt to estimate the direct cost of GDPR by studying the financial statements (specifically, profitability and SG&A expenses) of firms affected by the regulation. I con-

[4]In addition to these papers studying European firms, Liu (2024) studies the impact of GDPR on US listed firms, documenting that firms that are more exposed to GDPR (as proxied by the size of their legal departments) reduce data-processing activities after the implementation of GDPR.

tribute to this literature by providing direct evidence of the compliance costs of GDPR on firms.

I also contribute to the literature on impact of regulatory costs on young firms and on rates of entrepreneurship. Prior work, such as Peukert et al. (2022) and Goldberg et al. (2024) have documented that the impact of GDPR appear to disproportionately affect small firms while Trebbi, Zhang, and Simkovic (2026) note a similar pattern with respect to regulations in general. Jia et al. (2025) show that GDPR led to less trans-Atlantic VC investment to European startups while Mulla (2025) shows that AI regulation is associated with lower levels of VC investment into AI startups in the US. Methodologically, my paper is very similar to Trebbi et al. (2026) in that I use workforce data to estimate the costs of compliance. Relative to these papers, I am the first to document the significant costs of GDPR, a major European regulation, on firm growth and the number of new high-growth startups being founded.

Methodologically, this paper is similar to other recent papers using LinkedIn / Revelio data for cross-border studies of employment and startup activity. This provides good coverage compared to existing data sources, because a) cross-country surveys usually sample a very limited number of workers per country, b) administrative microdata provides high detail and good coverage, but is limited to single countries and c) aggregated national statistics are usually not detailed enough to distinguish between granular categories or workers or firms. LinkedIn data are particularly useful for studying white-collar work within high-growth startups, were coverage is likely to be highest. My paper builds on work by, for example, Amanzadeh, Kermani, and McQuade (2024) or Chen, Gao, Leung, and Wang (2025), as well as numerous practitioner/policy reports such as Atomico (2023) and Pal, Schneider, and Nurski (2025).

While regulations such as GDPR aim to protect consumer privacy, they may inadvertently disadvantage smaller firms relative to larger ones (Draghi, 2024). This paper does not aim to provide a thorough cost-benefit analysis of GDPR, but instead documents some of the costs of GDPR and their impact on small firms. By this, I aim to contribute to the growing

policy debate on the relative stagnation of Europe compared to the US, especially in the technology sector.

The remainder of the paper proceeds as follows. Section 2 provides institutional background on GDPR. Section 3 describes the data and sample construction. Section 4 presents the main results. Section 5 concludes.

2 Background

The General Data Protection Regulation (GDPR) is a comprehensive data privacy law enacted by the European Union. It was approved by the European Parliament in April 2016 and became enforceable on May 25, 2018. GDPR replaced the 1995 Data Protection Directive as well as a patchwork of national data protection laws and significantly expanded the rights of individuals regarding their personal data while imposing stricter obligations on organizations that collect, process, or store such data.

GDPR applies to any organization that processes personal data of EU residents, regardless of where the organization is located. This extraterritorial reach means that US-based companies serving EU customers must also comply with GDPR requirements. Key provisions include requirements for explicit consent before collecting personal data, the right to access and delete personal information (the "right to be forgotten"), mandatory data breach notifications within 72 hours, and appointment of Data Protection Officers for certain types of organizations.

Non-compliance can result in substantial penalties—up to 4% of annual global turnover or €20 million, whichever is higher. These penalties create particularly strong incentives for compliance among data-intensive firms that rely heavily on personal data for their core operations.

While national data privacy regulations already imposed some costs prior to the implementation of GDPR, GDPR is generally considered to have made data privacy enforcement

tougher across the EU (Johnson, 2022). In particular, some national regulations were looser than what GDPR required and it was common for companies to ignore them (Johnson, 2022). The penalties imposed by GDPR were also generally larger than those imposed by national regulators. While the European Commission's initial pre-legislation impact assessments (European Commission, 2012) argued that GDPR would reduce costs through harmonization of national privacy regulations, these estimates were seen as optimistic by for instance the UK (Ministry of Justice, 2012).

For startups, GDPR compliance imposes both direct and indirect costs. Direct costs include hiring specialized compliance personnel (Data Protection Officers, privacy lawyers, compliance managers), implementing technical safeguards (data encryption, access controls, audit trails), and conducting privacy impact assessments. Indirect costs arise from operational constraints, such as limits on data collection and use that may restrict product features or business models, and the opportunity cost of diverting engineering resources to compliance rather than product development.

3 Data

3.1 Revelio Labs Employment Data

The primary data source is Revelio Labs, a workforce intelligence platform that aggregates and structures publicly available professional profiles, primarily from LinkedIn. The dataset covers individual employment spells from 2010 to 2025, providing detailed information on job titles, employment dates, company identifiers, geographic locations, and worker characteristics. I use several tables from Revelio:

- The "positions"-table. This table contains information on employment spells, including the start and end date, role (Revelio's internal classification as well as an O*NET classification), location and company identifier

- The "company information"-table. This table contains information about companies, such as the headquarters location, a brief description (from their LinkedIn page), an industry classification as well as an estimate of the founding year. Revelio estimates these mainly from LinkedIn data and hence they are often unreliable for smaller firms
- The "worker education"-table. This table contains information on the educational history of each worker, including the institution, dates of study, the level of study and the field of study (if available)

Revelio's methodology involves scraping publicly visible LinkedIn profiles and applying machine learning algorithms to standardize and enrich the data. Each employment spell includes the worker's anonymized identifier, company identifier (allowing tracking of workers across firms), job title, start and end dates, and company characteristics including headquarters location, industry classification, and firm age. Revelio classifies job titles into approximately 17,000 distinct roles based on title text, work history patterns, and educational backgrounds. Additionally, Revelio maps job titles to O*NET occupation codes, enabling linkage to task-based occupation characteristics.

My unit of observation is at the employment-spell level, not at the firm-year level. Studying employment spells allows me to map *when* a position is created within a company, whereas an aggregation of the data to the firm-year (or firm-month) level would only tell me the coarse composition of a firm's workforce at any point in time. Since startups grow rapidly, understanding when (as a function of the firm's growth trajectory) a position in compliance is created can be informative. There are however drawbacks, such as the fact that employment spells in different firms (and regions) may last different lengths of time. I plan to add supplementary analyses using data aggregated to the firm-year level in later drafts.

The final dataset used in this analysis contains over 10 million employment spells across thousands of data-intensive startups (defined below) in the EU and US. The unit of observation is at the worker-position level, henceforth referred to as employment spells (a worker

may have multiple spells at the same firm if they change position).

3.2 Sample Construction

I am to identify high-growth data-intensive (HGDI) startups. Identifying high-growth firms is difficult without survey data. As I do not have survey data on the growth intentions of the firms in the Revelio Sample, I condition on the outcome variable, i.e. a firm achieving some measure of growth. I describe the steps taken to identify the firms for this sample below. I provide a more detailed guide to the construction of the sample (step-by-step) in Appendix A.1.

I start by loading the Revelio position-level data from 2010 to 2025. I drop all firms that have new employment spells in fewer than three years.

Step 1: Identify startup firms. I begin by identifying each company's "birth date" in two ways - using the "year founded" variable from Revelio as well as the date that the first employee joined the firm.[5] I drop firms whose founding date (according to Revelio) is prior to 2010.

I then calculate the number of employment spells per firm-year. I drop firms with more than 100 new spells in the first year to avoid picking up existing firms whose name or identity changes.[6]

To distinguish high-growth startups from small businesses in general, I calculate the differences between the highest annual number of new employment spells and the number of new employment spells in the first year and drop firms where this difference is lower than five. The goal of this is to drop firms that never gain any traction, as measured by the number of employees.

I also drop firms with names containing "stealth" since these refer to startups whose business idea is not known to the public yet (colloquially called "stealth startups") but get

[5]While these match in most cases, I use the date of the first employee joining as my preferred measure.

[6]While this may cause some *very* high growth startups (or e.g. "platform" startups such as Uber) to be mistakenly dropped, most startups, even ones that eventually become very successful, do not reach 100 employment spells in their first year.

incorrectly grouped as one company, as well as employment spells without an employer listed.

Step 2: Focus on data-intensive industries. GDPR compliance requirements are most binding for firms that rely heavily on personal data for core operations. To identify such firms, I use an AI-based classification of Revelio's 400-category industry taxonomy. Using Claude 4.5 Sonnet, I classify each industry as either being data intensive or not, and then classify its data intensity on a 1–10 scale based on its reliance on personal data, considering factors such as whether the industry involves direct consumer data collection, behavioral tracking, or personalized services.[7] Examples of industries classified as low and high data intensity can be found in Table 1.

I manually exclude public sector organizations, education, healthcare, nonprofits, hospitality, and commercial aviation, as these are either heavily regulated for other reasons, not-profit-motivated or unlikely to be startups.[8] I exclude legal services as this sector provides compliance services and hence the compliance share within the sector is very high, but unlikely to be representative of a regulatory burden on the firms themselves.

Step 3: Geographic coverage. I focus on firms headquartered in the European Union or the United States. I classify firms as "EU" or "US" based on headquarters location provided by Revelio. For the purposes of this analysis, I include the UK (as the UK left at the end of 2020 but remained compliant with many EU rules) and Norway as being part of the EU as both fully implemented GDPR in 2018, and include dependencies such as Reunion or the Faroe Islands as being part of the EU.[9]

Step 4: Sample restrictions. As my focus is startups, I focus on the first 300 employment spells within each firm, regardless of the length of these spells. Firms growing past 300 employment spells can be considered mature firms, beyond the scope of this analysis. My analysis sample period spans 2015-2021, i.e. within three years of GDPR's implementation. I define the analysis sample depending on the definition of the "treatment" variable (GDPR)

[7]The prompt used is in A.2.

[8]I drop these before excluding non-data-intensive sectors. It is very likely that many of these sectors would have been eliminated at the next step anyway.

[9]Given the scarcity of startups in these areas, it makes no difference whether I include or exclude them.

- for instance, if the "treatment" status of a firm is based on its founding date, I look at firms founded between 2015 and 2021. If treatment status is defined based on the start date of an employment spell, I look at employment spells starting between 2015 and 2021.

The final sample consists of almost 90,000 data-intensive startups (founded between 2010 and 2025) and about 12 million employment spells. Of these, about 38 thousand startups were "founded" (based on the date of the first employee) between 2015 and 2021 and with 6.5 million employment spells starting between those years. These startups and workers make up my main analysis sample.

3.3 Measuring Compliance Costs

I employ two complementary measures of regulatory compliance costs.

Measure 1: AI-classified compliance roles (main measure). I classify employment spells into compliance-related roles using the 17,000-category job title taxonomy from Revelio Labs.[10] Using Claude 4.5 Opus, I categorize each of the 17,000 roles into one of three categories: (1) *Compliance* roles include privacy officers, data protection officers, legal counsel specializing in privacy, compliance managers, and regulatory affairs specialists; (2) *Core Operations* roles include software engineers, data scientists, product managers, sales personnel, and marketing staff—positions directly involved in product development and revenue generation; (3) *Other* roles include finance, HR, administration, and general support functions. The prompt used to classify roles is provided in Appendix A.2.

Measure 2: Regulatory Tasks Index. Following Trebbi et al. (2026), I construct an occupation-level measure of regulatory intensity using O*NET task descriptions. The Regulatory Tasks Index (RTI) quantifies the extent to which an occupation involves tasks related to regulatory compliance, legal requirements, and policy adherence. Essentially, the measure uses data on the tasks associated with various job titles, and classifies certain tasks as being related to regulatory compliance. Then, for each job title, an index is created based

[10]Revelio's 17,000 role taxonomy is their latest classification and most up-to-date, but the results are robust to using their previous 1,500 role classifcation.

on what percentage of the tasks associated with that role are related to compliance. The measure has several drawbacks in my setting: first, it is based on US data. Tasks may not be divided in the same way in Europe. Second, it is based on all firms across the economy. In startups, early employees may take on tasks that they would not at larger firms.

While Trebbi et al. (2026) make their data publicly available, they are based on an older version of the O*NET classification (23). I use Claude 4.5 Sonnet to replicate the methodology of Trebbi et al. (2026) for O*NET v30, i.e. first searching task descriptions for regulation-related keywords and then classifying these descriptions as either true or false positives. One key difference is that v30 of the O*NET database does not have an importance metric for all tasks, so I treat all tasks as being equal (unlike Trebbi et al. (2026), who weight tasks based on their importance). As I merge RTI scores to individual employment spells using Revelio's O*NET occupation codes.[11]

Both measures likely understate the true compliance burden of GDPR. Many "core" workers perform compliance-related tasks without being classified as compliance specialists: for example, CEOs managing GDPR compliance, engineers implementing data minimization, or product managers redesigning features to reduce data collection. Thus, estimated effects should be interpreted as lower bounds on the total compliance burden.

3.4 Data Limitations

Several limitations should be noted. First, the data only include workers with LinkedIn profiles, which may underrepresent certain demographics or industries. However, LinkedIn penetration is high among white-collar workers in technology and business services, especially knowledge-intensive startups. Core et al. (2024) merge Revelio data with data on VC-backed startups from LSEG Workspace (formerly VentureXpert) and Pitchbook and find roughly 75% of VC-backed startups in the Revelio data. This is likely to understate the representation

[11]The updated classification can be downloaded from: https://github.com/mpaaso/GDPR-and-Firm-Growth/tree/main/AI%20Classified%20ONET%20V30. Please cite Trebbi et al. (2026) even if using my classification.

of these firms, as startups may change names, have a similar name to other startups, get acquired or something similar, leading to matches being missed. In terms of threats to identification, any systematic differences in LinkedIn usage between EU and US should be captured by the difference-in-differences design.

Second, employment dates are self-reported and may contain errors or omissions. I attempt to mitigate this by, for example, dropping individuals who join a company well before the second employee and are located in a different country. There may also be people embellishing their LinkedIn profiles by entering fake employers, but this is less likely to be an issue for startup firms.

Third, LinkedIn usage has increased over time. This means that measures such as the likelihood that a firm reaches a threshold number of employees are not comparable across time by themselves, as a higher proportion of firms may appear to reach thresholds simply because more workers are on LinkedIn. Because of this, the bulk of the comparisons I make in the paper are comparisons of EU- and US-headquartered firms. It is also likely that there are differences in LinkedIn usage between the EU and the US. However, since my comparison is both across regions and across time, these differences would need to be time-variant to affect my results (as well as to disproportionately affect data-intensive startups).

Finally, as noted above, job titles are very unlikely to capture the full impact of regulatory burdens on startups. Job titles in young firms can be very fluid, and founders and early workers are very likely to be working on tasks that their job titles (or the RegIndex) do not fully capture. This means that my estimates are always likely to understate the regulatory burden that startups faced.

Here it is important to note that a concern with simply comparing US to European firms could be that in one of the two markets, compliance tasks may be outsourced to third parties (such as GDPR consultants) moreso than in the other. Because of this concern, my main results are based on a quasi-differences-in-differences analysis, where I compare the *change* in regulatory employees and tasks around the implementation of GDPR rather than simply

comparing US startups with European ones.

3.5 Summary Statistics

Table 2 presents summary statistics for the analysis sample, which consists of employment spells at firms classified as data intensive (AI response was "Yes"). These firms have data intensity scores ranging from 6 to 10. Panel A presents individual-level data whereas Panel B presents firm-level data. There are 9.9 million employment spells in my sample (i.e. spells at data-intensive startups, only the first 300 spells, spells starting between 2015 and 2021). These spells are roughly evenly split between the US and the EU. EU workers are more likely to have compliance-related roles, though Trebbi et al. (2026)'s RTI is similar across the two regions.

There are 88,617 startups in my sample, with an average of 30 employment spells per startup. 15% of startups reach 50 employment-spells within four years of founding, 2.5% reach 150 and less than 1% reach 300 employment spells.

As an additional falsification test, I also study non-data-intensive firms. These firms are not included in the summary statistics table.

4 Results

4.1 Compliance Employees

To examine how GDPR affected the composition of labor within firms, I study job positions around the announcement and implementation of GDPR. My main outcome of interest is a dummy that takes the value of 1 if a position is classified as being related to compliance (as classified by Claude 4.5 Sonnet, see above). I also study whether GDPR affected the number of compliance tasks performed by a firm's workers, using the Regulatory Tasks Index (RTI) of Trebbi et al. (2026).

As noted above, the analysis focuses on the first 300 employment spells by data-intensive-startups and on job spells starting between 2015 and 2021 (tests based on GDPR implementation in 2018) or spells starting between 2013 and 2019 (tests based on GDPR's announcement in 2016).

In Figure 1, I show visually the share of workers in compliance roles in EU and US companies, sorted by the starting date of the role. I plot the share from 2010 to 2025 despite the analysis sample being shorter to illustrate that in the past, the share of compliance workers in the US and the EU moved in parallel until 2016, when GDPR was announced.

I also plot the share of compliance workers at non-data-intensive startups, which has also moved roughly in parallel in the US and EU until about 2016. Note that the compliance shares are not comparable and that these images should not be interpreted as saying that non-data-intensive startups hire more compliance staff - it may simply be that a larger share of normal "production" employees at data-intensive startups use LinkedIn.

There is a large increase in the share of EU employees in compliance roles starting in 2016 until 2018. GDPR was first announced in 2016, suggesting that firms prepared for its implementation at least to some extent by hiring compliance workers.

I then move to a formal analysis.

I estimate the following quasi-difference-in-differences specification:

$$Y_{ijt} = \beta_1(\text{EU}_j \times \text{Post-GDPR}_t) + \gamma_t + \delta_j + \lambda_i + \varepsilon_{ijt} \tag{1}$$

where Y_{ijt} denotes the outcome (compliance-role or RTI) for employment spell i at firm j with start year t. The variable EU_j indicates whether the firm is headquartered in the European Union, and Post-GDPR_t equals one for positions starting after 2016 or 2018, depending on whether the specification is based on GDPR's announcement or implementation date. The coefficient of interest, β_1, captures the differential change in worker outcomes at EU firms relative to US firms following GDPR implementation. I include year ($gamma_t$) and firm fixed-effects (δ_j) in all specifications and employment-spell-number fixed-effects (λ_i) in

columns (2), (4), (6) and (8) to control for the order in which an employee joins a firm. The variation I am exploiting therefore comes from within-firm changes at EU-headquartered firms after 2018, controlling for broader trends in the share of compliance employees.

Table 3 presents results. In Panel A, outcome variable is the compliance-role dummy whereas the outcome variable in Panel B is the Regulatory Tasks Index of Trebbi et al. (2026). $PostGDPR$ in olumns (1)–(4) of both tables is based on the implementation year of GDPR (2018) whereas in columns (5)–(8) of both tables, it is based on the announcement year of GDPR (2016).

In columns (3), (4), (7) and (8) of both panels, I restrict the sample to the most data-intensive industries, i.e. those with a data intensity score of 9 or 10.

The EU x Post-GDPR interaction is positively associated with the probability of a position being compliance-related. In terms of economic significance, the 0.11 percentage-point increase in the probability of a role being compliance related (Panel A, column (3)) for the most data-intense firms is roughly 20% of the gap between the unconditional probability of an EU hire being in compliance vs a US one (which is about 0.6 percentage points).

In the top panel of Figure 2, I provide "event study"-style dynamic differences-in-differences estimates of compliance share in EU-headquartered startups vs. US-headquartered ones. There are no differences between the two until about 2017, when the compliance share in EU startups starts to rise. The bottom panel of Figure 2 presents regression estimates from Table 3 with alternative treatment years. For instance, the estimate for 2015 treats 2015 as the year in which GDPR was implemented, and compares compliance share in the EU vs. the US between 2012 and 2018, with 2016-2018 being considered "treated". The EU x Post coefficient is insignificant until 2015, when most of the treatment period is actually after the announcement of GDPR, and remains positive and significant until 2020, when both the pre- and post-treatment periods occur after the implementation of GDPR. In 2022, the coefficient is positive and significant again, reflecting the rise in compliance share in the EU after 2021 seen in Figure 1 (which is likely driven by other EU regulations enacted around

the time).

With regards to regulatory tasks, the coefficient is positive in all columns, but rarely statistically or economically significant. On average, a position has an RTI of 0.023, or in other words, its weighted-average regulatory-task intensity is 2.4%. The estimate in column 2 implies that this increased by roughly 0.0002 index points after the implementation of GDPR, an increase of less than 1%.

4.1.1 Non-Data-Intensive Firms and Data Intensity Interaction

I repeat these regressions for non-data-intensive firms. The results are presented in Table 4. While graphically (Figure 1) GDPR did not seem to be associated with an increase in the compliance share of non-data-intensive startups, the coefficients in Table 4 indicate that the probability of a job spell being related to compliance rose in these firms as well. This may be driven by the fact that in Figure 1, compliance share in non-data-intensive startups is falling sharply, especially in the EU, prior to the implementation of GDPR.

Nevertheless, it is plausible that GDPR also affected non-data-intensive startups, as for instance almost all firms maintain some databases with customer data, which would be affected by the change.

I then formally test whether the increase in compliance share was higher for HGDIs than non-data-intensive startups. I do this by running a triple-interaction of EU, $PostGDPR$ and $DIScore$, the data-intensity score (ranging from 0–10) assigned by the LLM.

The results are presented in Table 5. In columns (1) and (2), $PostGDPR$ equals one after 2018 whereas in columns (3) and (4), $PostGDPR$ equals one after 2016. The $DIScore \times EU \times PostGDPR$ interaction is positive and significant when based on the GDPR announcement date, but not based on the effective date. As noted above though, this test may be biased downwards because of the (presumably unrelated to GDPR) downward trend of compliance share in non-data-intensive EU startups prior to 2016.

4.2 Startup Growth

I then test whether GDPR affected the growth rates of HGDI startups. I start by plotting the rates at which HGDI firms reach certain employment thresholds within four years of founding (once again, driven by the fact that my sample ends in 2021 and my data in 2025). Figure 3 presents the percentage of HGDI startups founded in any given year that reach an employment threshold.

Prior to GDPR's announcement in 2016, there are no significant differences between the likelihood of an EU and US startup reaching any employment threshold (other than perhaps 50 employment spells). However, after GDPR was announced and especially after it was implemented, US firms are more likely to reach every threshold.

In Figure 4 I repeat this analysis for non-data-intensive startups and find no differences in the likelihood of reaching any threshold other than 50 employees after the implementation of GDPR in 2018, but this gap is much about half as big as the gap for HGDI startups.

Please note that in the graphs above, the rising share of firms reaching various employment thresholds is likely driven by the increasing share of the workforce using LinkedIn. Hence it is important to compare European vs. US startups separately.

I test this formally in two ways: First, I test whether firms *founded after GDPR* was announced/implemented in the EU are more or less likely to reach certain hiring thresholds within four years. Second, I test whether startups in the EU are less likely to reach certain hiring thresholds after the implementation of GDPR. Essentially, the first test is a cross-sectional analysis whereas the second is based on firm-year data. The second specification tests, conditional on reaching an employment threshold, how fast firms reach it.[12]

I estimate the following models:

$$Y_i = \beta_1 \text{EU}_i + \beta_2(\text{EU}_i \times \text{Founded Post-GDPR}_i) + \gamma_t + \delta_s + \varepsilon_{it}$$

[12]The firm fixed-effect means that firms that never reach a threshold are effectively omitted from this analysis.

Where Y_i is a dummy that takes the value of 1 if a startup has reached a certain hiring threshold (50, 150, 200 or 300 hires) within four years of its founding, EU_i is a dummy take equals 1 if the startup is headquartered in the EU and $FoundedPostGDPR_i$ is a dummy that takes the value of 1 if the firm was founded after 2016 (GDPR announced) or 2018 (GDPR effective). γ_t represents year-of-founding fixed-effects while δ_s represents RICS-k400-industry fixed-effects.

And:

$$Y_{it} = \beta_1(\text{EU}_i \times \text{Post-GDPR}_t) + +\gamma_t + \delta_i + \varepsilon_{it}$$

Where Y_{it} is a dummy that takes the value of 1 once a startup has reached a certain hiring threshold (50, 150, 200 or 300 hires), EU_i is a dummy as above and Post-GDPR_t is a dummy that equals 1 after 2016/2018 (depending on the specifications). I include firm and year fixed-effects.

The results are presented in Table 6. Panel A presents the cross-sectional analyses whereas Panel B presents the within-firm analysis. The $\text{EU}_i \times \text{Founded Post-GDPR}_i$ dummy is negative in all specifications. The coefficient in Panel A, Column (1), implies that post-GDPR, startups headquartered in the EU were 2.6 percentage points less likely to reach 50 employment spells after GDPR's implementation in 2018, compared to a baseline pre-GDPR rate of roughly 15% and a pre-GDPR US-EU gap of about 1 percentage point. The effects are relatively larger for higher thresholds, where the unconditional probability of reaching these thresholds is already lower.

In Panel B, the $\text{EU}_i \times \text{Post-GDPR}_t$ coefficient is negative and significant for all employment thresholds. The coefficient in column (1) implies that in any given year after GDPR, an EU-headquartered firm is 0.9 percentage points less likely to reach 50 employment spells. As the firm fixed-effects mean that no variation is coming from firms that never reach the threshold number of employment spells, this is essentially a test of the speed at which firms reach employment spells.

4.2.1 Non-Data-Intensive Firms - Startup Growth

Once again, I repeat the analysis with non-data-intensive startups. The results are presented in Table 7. Panel A suggests no significant differences in the probability of EU and US firms reaching employment thresholds within four years after GDPR. This is consistent with Figure 4. In Panel B, it does appear that in any given year, a European startup is less likely to reach employment thresholds, but this effect is less than half of the effect size for HGDIs for the thresholds above 50 employment spells.

Finally, I run a triple-interaction with the data-intensity score assigned by Claude 4.5 Sonnet (ranging from 0–10, though in practice the lowest score is 2). The results are presented in Table 8. For most thresholds, and for both the firm-level and position-level analysis, the interaction $DIScore \times EU \times PostGDPR$ is negative and significant, suggesting that data-intensive firms were less likely to achieve high-growth after the implementation of GDPR as well as the announcement of GDPR (though less so).

5 Conclusion

The implementation of GDPR represents one of the most significant recent European regulations, and its footprint is clearly visible in the reconfiguration of the labor market. I show that the regulation leads to higher spending on compliance-roles at European startups, and is associated with lower rates of entrepreneurship and startup growth.

My estimates imply that GDPR explains roughly 10–20% of the post-2018 "compliance gap" between the US and the EU. Unconditionally, this gap means that startups in the EU hire on average 20% more compliance staff than corresponding US ones. I do not find strong evidence of an increase in regulatory tasks using the Regulatory Tasks Index of Trebbi et al. (2026), though this may be driven by its unsuitability for startups and cross-country comparisons.

As data privacy regulations continue to expand globally, my findings suggest that there

is a real cost of these regulations for startups and innovation.

References

Amanzadeh, Naser, Amir Kermani, and Timothy McQuade, 2024, Return Migration and Human Capital Flows .

Atomico, 2023, The state of european tech, Industry report, Atomico.

Bombardini, Matilde, Francesco Trebbi, and Miao Ben Zhang, 2024, Measuring the costs and benefits of regulation, *Annual Review of Economics* 17.

Bradford, Anu, 2020, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press, New York).

Chen, Lei, Chaoyue Gao, Alvin Leung, and Xiaoning Wang, 2025, When Assurance Undermines Intelligence: The Efficiency Costs of Data Governance in AI-Enabled Labor Markets, arXiv:2511.01923 [cs].

Chen, Yuchen, Can Huang, and Xuelai Li, 2026, Technology Literacy and Deep-Tech Investment: Evidence from VC Industry .

Core, Fabrizio, Gosia Ryduchowska, and Su Wang, 2024, Human capital acquisition in vc-backed companies, *Available at SSRN* .

Deisenroth, Daniel, Utsav Manjeer, Zarak Sohail, Steven Tadelis, and Nils Wernerfelt, 2024, Digital Advertising and Market Structure: Implications for Privacy Regulation, NBER Working Paper, National Bureau of Economic Research.

Demirer, Mert, Diego Jiménez Hernández, Dean Li, and Sida Peng, 2024, Data, Privacy Laws and Firm Production: Evidence from the GDPR, NBER Working Paper, National Bureau of Economic Research.

Draghi, Mario, 2024, The future of european competitiveness: A competitiveness strategy for europe.

European Commission, 2012, Commission staff working document: Impact assessment accompanying the document regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) and Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data, Technical Report SEC(2012) 72 final, European Commission, Brussels.

Frey, Carl Benedikt, and Giorgio Presidente, 2024, Privacy regulation and firm performance: Estimating the GDPR effect globally, *Economic Inquiry* 62, 1074–1089.

Garicano, Luis, Bengt Holmström, and Nicolas Petit, 2025, The Constitution of Innovation: A New European Renaissance .

Goldberg, Samuel G., Garrett A. Johnson, and Scott K. Shriver, 2024, Regulating Privacy Online: An Economic Evaluation of the GDPR, *American Economic Journal: Economic Policy* 16, 325–358.

Janßen, Rebecca, Reinhold Kesler, Michael E. Kummer, and Joel Waldfogel, 2022, GDPR and the Lost Generation of Innovative Apps, Working Paper, National Bureau of Economic Research.

Jia, Jian, Ginger Zhe Jin, Mario Leccese, and Liad Wagman, 2025, How Does Privacy Regulation Affect Transatlantic Venture Investment? Evidence from GDPR, NBER Working Paper, National Bureau of Economic Research.

Johnson, Garrett, 2022, Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond, NBER Working Paper, National Bureau of Economic Research.

Johnson, Garrett A., Scott K. Shriver, and Samuel G. Goldberg, 2023, Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR, *Management Science* 69, 5695–5721, Publisher: INFORMS.

Koski, Heli, and Nelli Valmari, 2020, Short-term impacts of the gdpr on firm performance, Technical report, ETLA Working Papers.

Liu, Po-Yu, 2024, Value of Data to U.S. Public Firms: Evidence from GDPR .

Ministry of Justice, 2012, Impact assessment: Proposal for an EU data protection regulation, Technical report, Ministry of Justice, London, UK.

Mulla, Junida, 2025, AI Regulation, Startup Financing, and Talent Allocation .

Pal, Siddhi, Catherine Schneider, and Laura Nurski, 2025, Solving europe's ai talent equation: Supply, demand and missing pieces, Policy report / data brief, Centre for European Policy Studies (CEPS) and Interface, Uses labour market data from Revelio Labs and Lightcast for cross–country AI talent and vacancy analysis.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer, 2022, Regulatory Spillovers and Data Governance: Evidence from the GDPR, *Marketing Science* Publisher: INFORMS.

Trebbi, Francesco, Miao Ben Zhang, and Michael Simkovic, 2026, The Cost of Regulatory Compliance in the United States, *Review of Financial Studies* Forthcoming.

Table 1
Industry Data Intensity Classification

This table presents the AI-based classification of Revelio's 400-category industry taxonomy by data intensity. Industries were classified on a 1–10 scale based on their reliance on personal data for core operations using Claude 4.5 Sonnet. See Appendix A.2 for the prompt used. Panel A shows industries with the highest data intensity (score = 10) Panel B shows the first 25 (alphabetically sorted) industries with high data intensity (score = 9). Panel C shows the 25 industries with the lowest data intensity scores, with the score for each industry in brackets.

Panel A: Highest (Score=10)	**Panel B: High (Score=9)**	**Panel C: Lowest (Bottom 25)**
Digital Marketing and Advertising	Artificial Intelligence and Software Development	Building Materials and Interior Finishing Services (2)
Digital Marketing and Analytics	Banking	Cement and Concrete Manufacturing (2)
Healthcare Information Technology	Banking and Financial Services	Confectionery and Specialty Food Ingredients Producers (2)
Tech and E-Commerce Platforms	Business Intelligence and Analytics	Creative and Design Services (2)
	Digital Health and Wellness Services	Food Processing and Production (2)
	Digital Marketing and Communications	Food and Beverage Processing (2)
	Digital Media and Entertainment	Furniture and Home Decor Manufacturing (2)
	Digital Payment Platforms	Graphic and Industrial Design Services (2)
	Digital Services and E-Commerce	Home Appliances and Building Materials (2)
	E-Commerce and Technology Services	Metalworking and Machinery Manufacturing (2)
	Employment and Recruitment Services	Mining and Metals (2)
	Financial Markets and Analytics Services	Packaging and Labeling Solutions (2)
	Financial Services	Packaging and Paper Products (2)
	Financial Services and Asset Management	Printing and Graphic Communications (2)
	Financial Services and Capital Markets	Printing and Packaging (2)
	Financial Services and Credit Cooperatives	Pulp, Paper, and Chemicals (2)
	Financial Services and Credit Unions	Steel and Metals Manufacturing (2)
	Financial Services and Fintech	Agricultural Inputs and Services (3)
	Financial Services and Insurance	Agricultural Products and Services (3)
	Financial Services and Investment Banking	Agrochemicals and Agricultural Sciences (3)
	Financial Services and Investment Management	Alcoholic Beverages and Brewing (3)
	Financial Services and Microfinance	Architectural and Engineering Design Services (3)
	Financial Technology and Services	Architectural and Urban Design Services (3)
	Financial and Professional IT Services	Arts and Culture (3)
	Fintech and Online Marketplaces	Automation and Robotics Equipment (3)

Table 2
Summary Statistics

Summary statistics for high-growth data-intensive startups in the EU and US, 2015–2021. The sample includes firms with data intensity score > 8 and the first 300 employment spells per firm. Panel A presents summary statistics for key variables at the job-position level, whereas Panel B presents summary statistics at the firm-level.

Panel A: Position-Level Variables

	(1) All			(2) EU			(3) US		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Indicator: Compliance role	9,999,820	0.017	0.129	4,581,006	0.020	0.142	5,418,814	0.014	0.117
Trebbi et al: Regulation-Intensity Index	9,392,094	0.024	0.056	4,299,045	0.025	0.058	5,093,049	0.024	0.055
Employee order within the company	9,999,893	81.369	75.821	4,581,049	76.674	73.895	5,418,844	85.339	77.189
Observations	9999893			4581049			5418844		

Panel B: Firm-Level Variables

	(1) All			(2) EU			(3) US		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total employment spells	88,617	30.377	39.728	44,387	29.900	37.962	44,230	30.855	41.420
Reached 50 employment spells	88,617	0.154	0.361	44,387	0.150	0.357	44,230	0.159	0.365
Reached 150 employment spells	88,617	0.023	0.150	44,387	0.021	0.142	44,230	0.026	0.158
Reached 300 employment spells	88,617	0.005	0.071	44,387	0.004	0.066	44,230	0.006	0.076
Observations	88617			44387			44230		

Table 3
GDPR and Worker Compliance Roles and Tasks

This table reports regressions where the outcome variable is a position-level measure of compliance intensity. The sample consists of the first 300 employment spells at data-intensive startups between 2015 and 2021 in Columns (1)–(4) (GDPR implementation date) and 2013 to 2019 in Columns (5)–(8) (GDPR announcement date). **Panel A** presents the results of linear probability models where the dependent variable is an indicator for whether the worker holds a compliance role. **Panel B** presents the results of OLS regressions where the dependent variable is the Regulatory Task Index (RTI), a continuous measure of regulatory task intensity associated with each position. The *EU* variable equals 1 if a startup is headquartered in the EU. $Post-GDPR$ equals 1 if an employment spell started after 2018 in columns (1)–(4) and after 2016 in columns (5)-(8). In Columns (3), (4), (7) and (8), I focus on the most data intensive firms - those with a data intensity score of 9 or 10. All columns include firm and year fixed-effects. Columns (2), (4), (6) and (8) add employment-spell number (e.g. indicating the 10th hire at a firm) fixed effects. Standard errors are clustered at the firm level.

Panel A: Compliance Role Dummy

	GDPR Effective (2018)				GDPR Announcement (2016)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EU × Post-GDPR Effective	0.0005**	0.0006***	0.0011***	0.0012***				
	(0.0002)	(0.0002)	(0.0003)	(0.0003)				
EU × Post-GDPR Announcement					0.0008***	0.0009***	0.0021***	0.0022***
					(0.0003)	(0.0003)	(0.0004)	(0.0004)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Emp Spell Num FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	9,986,030	9,986,030	4,315,168	4,315,168	7,250,472	7,250,472	3,156,036	3,156,036
Adjusted R^2	0.141	0.141	0.078	0.078	0.142	0.142	0.080	0.080

Panel B: RTI

	GDPR Effective (2018)				GDPR Announcement (2016)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EU × Post-GDPR Effective	0.0001	0.0002**	0.0001	0.0002				
	(0.0001)	(0.0001)	(0.0002)	(0.0002)				
EU × Post-GDPR Announcement					0.0002**	0.0004***	0.0002	0.0004**
					(0.0001)	(0.0001)	(0.0002)	(0.0002)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Emp Spell Num FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	9,378,687	9,378,687	4,058,558	4,058,558	6,812,791	6,812,791	2,970,045	2,970,045
Adjusted R^2	0.151	0.152	0.111	0.112	0.155	0.156	0.116	0.117

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4
GDPR and Worker Compliance Roles and Tasks - Non-Data-Intensive

This table reports regressions where the outcome variable is a position-level measure of compliance intensity. The sample consists of the first 300 employment spells at non-data-intensive startups between 2015 and 2021 (Columns (1), (2), (5) and (6), based on GDPR's implementation date) and 2013 to 2019 (Columns (3), (4), (7) and (8), based on GDPR's announcement date). Columns (1)–(4) estimate linear probability models where the dependent variable is an indicator for whether the worker holds a compliance role. Columns (5)–(8) estimate OLS regressions where the dependent variable is the Regulatory Task Index (RTI), a continuous measure of regulatory task intensity associated with each position. The *EU* variable takes the value of 1 if a startup is headquartered in the EU. $Post-GDPR$ takes the value of 1 is an employment spell started after 2018 in columns (1), (2), (5) and (6) and after 2016 in columns (3), (4), (7) and (8). All columns include firm and year fixed-effects. Columns (2) and (4) add employment-spell number (e.g. indicating the 10th hire at a firm) fixed effects. Standard errors are clustered at the firm level.

	Compliance Role				Regulatory Task Index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EU × Post-GDPR Effective	0.0007**	0.0008**			0.0001	0.0002		
	(0.0003)	(0.0003)			(0.0001)	(0.0001)		
EU × Post-GDPR Announcement			0.0009**	0.0010**			0.0005***	0.0006***
			(0.0004)	(0.0004)			(0.0002)	(0.0002)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Emp Spell Num FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	5,826,497	5,826,497	4,383,511	4,383,511	5,424,718	5,424,718	4,068,800	4,068,800
Adjusted R^2	0.255	0.255	0.273	0.273	0.209	0.209	0.218	0.219

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5
GDPR and Worker Compliance Roles and Tasks - Interaction

This table reports regressions where the outcome variable is a position-level measure of compliance intensity. *DIScore* is a variable measuring the "data intensity" of a startup's industry. The *EU* variable takes the value of 1 if a startup is headquartered in the EU. $Post-GDPR$ takes the value of 1 is an employment spell started after 2018 in columns (1) and (2) and 2016 in columns (3) and (4). All columns include firm and year fixed-effects. Columns (2) and (4) add employment-spell number (e.g. indicating the 10th hire at a firm) fixed effects. Standard errors are clustered at the firm level.

	GDPR Effective (2018)		GDPR Announcement (2016)	
	(1)	(2)	(3)	(4)
DI Score × EU × Post-GDPR	0.0000	0.0001		
	(0.0001)	(0.0001)		
DI Score × EU × Post-GDPR			0.0002**	0.0002**
			(0.0001)	(0.0001)
EU × Post-GDPR	0.0003	0.0003		
	(0.0005)	(0.0005)		
EU × Post-GDPR			-0.0004	-0.0004
			(0.0006)	(0.0006)
DI Score × Post-GDPR	0.0002***	0.0002***		
	(0.0001)	(0.0001)		
DI Score × Post-GDPR			0.0003***	0.0003***
			(0.0001)	(0.0001)
Year FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Emp Spell Num FE	No	Yes	No	Yes
Observations	15812527	15812527	11633983	11633983
Adjusted R^2	0.195	0.195	0.206	0.206

Table 6
GDPR and Firm Growth

Panel A presents results of regression where the outcome variable is a dummy indicating whether a firm reaches a certain threshold of employment spells (e.g. 50 in column 1) *within the first four years of its operation*. *EU x Post-GDPR Announced/Effective* denotes the interaction of a dummy that equals 1 if the startup is headquartered in the EU and a dummy that equals one if the startup was founded after 2016 (Announced) or 2018 (Effective). The unit of observation is a firm. The table includes founding-year fixed-effects and sector (RICS k400) fixed-effects. Standard errors are clustered by sector.

Panel B presents results of regressions where the outcome variable is a dummy variable indicating that a firm has passed a threshold of employment spells. *EU x Post-GDPR Announced/Effective* is an interaction of a dummy variable that equals 1 if the startup is headquartered in the EU and a dummy variable that equals 1 if the year is after 2016 (Announced) or 2018 (Effective). All regressions include calendar-year and firm fixed-effects. Standard errors are clustered by firm.

Panel A: Cross-Section (Founding Year)

	(1) 50	(2) 150	(3) 300	(4) 50	(5) 150	(6) 300
EU × Post-GDPR Effective	-0.026***	-0.005***	-0.001			
	(0.006)	(0.002)	(0.001)			
EU × Post-GDPR Announcement				-0.007*	-0.004**	-0.001*
				(0.004)	(0.002)	(0.001)
Founding Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	No	No	No
Observations	88,617	88,617	88,617	103,369	103,369	103,369
Adjusted R^2	0.040	0.010	0.004	0.039	0.009	0.003

Panel B: Firm-Year Panel (Observation Year)

	(1) 50	(2) 150	(3) 300	(4) 50	(5) 150	(6) 300
EU × Post-GDPR Effective	-0.009***	-0.020***	-0.011***			
	(0.002)	(0.001)	(0.001)			
EU × Post-GDPR Announcement				-0.017***	-0.014***	-0.006***
				(0.002)	(0.001)	(0.001)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	No	No	No	No	No	No
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	962,227	962,227	962,227	778,240	778,240	778,240
Adjusted R^2	0.648	0.617	0.590	0.595	0.545	0.502

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7
GDPR and Firm Growth - Non-Data-Intensive

Panel A presents results of regression where the outcome variable is a dummy indicating whether a firm reaches a certain threshold of employment spells (e.g. 50 in column 1) *within the first four years of its operation*. *EU x Post-GDPR Announced/Effective* denotes the interaction of a dummy that equals 1 if the startup is headquartered in the EU and a dummy that equals one if the startup was founded after 2016 (Announced) or 2018 (Effective). The unit of observation is a firm. The table includes founding-year fixed-effects and sector (RICS k400) fixed-effects. Standard errors are clustered by sector.

Panel B presents results of regressions where the outcome variable is a dummy variable indicating that a firm has passed a threshold of employment spells. *EU x Post-GDPR Announced/Effective* is an interaction of a dummy variable that equals 1 if the startup is headquartered in the EU and a dummy variable that equals 1 if the year is after 2016 (Announced) or 2018 (Effective). All regressions include calendar-year and firm fixed-effects. Standard errors are clustered by firm.

Panel A: Cross-Section (Founding Year)

	(1) 50	(2) 150	(3) 300	(4) 50	(5) 150	(6) 300
EU × Post-GDPR Effective	-0.009	0.001	0.001			
	(0.006)	(0.003)	(0.001)			
EU × Post-GDPR Announcement				0.001	0.001	0.000
				(0.004)	(0.002)	(0.001)
Founding Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	No	No	No
Observations	49,964	49,964	49,964	61,941	61,941	61,941
Adjusted R^2	0.043	0.013	0.006	0.037	0.010	0.005

Panel B: Firm-Year Panel (Observation Year)

	(1) 50	(2) 150	(3) 300	(4) 50	(5) 150	(6) 300
EU × Post-GDPR Effective	-0.007***	-0.006***	-0.004***			
	(0.002)	(0.001)	(0.001)			
EU × Post-GDPR Announcement				-0.013***	-0.006***	-0.002*
				(0.002)	(0.002)	(0.001)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	No	No	No	No	No	No
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	640,469	640,469	640,469	527,702	527,702	527,702
Adjusted R^2	0.650	0.630	0.609	0.596	0.555	0.522

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8
GDPR and Firm Growth - Interaction

Panel A presents results of regression where the outcome variable is a dummy indicating whether a firm reaches a certain threshold of employment spells (e.g. 50 in column 1) *within the first four years of its operation. DI Score x EU x Founded Post-GDPR* denotes the interaction of the data-intensity score (assigned by Claude, ranging from 0–10), a dummy that equals 1 if the startup is headquartered in the EU and a dummy that equals one if the startup was founded after 2016 (Announced) or 2018 (Effective). The unit of observation is a firm. The table includes founding-year fixed-effects and sector (RICS k400) fixed-effects. Standard errors are clustered by sector.

Panel A: Cross-Section (Founding Year)

	(1) 50	(2) 150	(3) 300	(4) 50	(5) 150	(6) 300
DI Score × EU × Post-GDPR Effective	-0.0040**	-0.0014**	-0.0005**			
	(0.0018)	(0.0007)	(0.0003)			
DI Score × EU × Post-GDPR Announcement				-0.0017	-0.0011**	-0.0001
				(0.0015)	(0.0005)	(0.0002)
EU × Post-GDPR Effective	0.0063	0.0064	0.0032*			
	(0.0118)	(0.0047)	(0.0017)			
EU × Post-GDPR Announcement				0.0058	0.0049	-0.0002
				(0.0101)	(0.0031)	(0.0013)
DI Score × Post-GDPR Effective	0.0027*	0.0010*	0.0005**			
	(0.0015)	(0.0006)	(0.0002)			
DI Score × Post-GDPR Announcement				0.0036***	0.0010**	0.0001
				(0.0012)	(0.0004)	(0.0002)
Founding Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	138,581	138,581	138,581	165,310	165,310	165,310
Adjusted R^2	0.042	0.011	0.004	0.041	0.010	0.003

Table 8
GDPR and Firm Growth - Interaction (continued)

Panel B presents results of regressions where the outcome variable is a dummy variable indicating that a firm has passed a threshold of employment spells. *DI Score x EU x Year is Post-GDPR* is an interaction of the data-intensity score (assigned by Claude, ranging from 0–10), a dummy variable that equals 1 if the startup is headquartered in the EU and a dummy variable that equals 1 if the year is after 2016/2018. All regressions include calendar-year and firm fixed-effects. Standard errors are clustered by firm.

Panel B: Firm-Year Panel (Observation Year)

	(1) 50	(2) 150	(3) 300	(4) 50	(5) 150	(6) 300
DI Score × EU × Post-GDPR Effective	0.0008	-0.0020***	-0.0011***			
	(0.0006)	(0.0004)	(0.0003)			
DI Score × EU × Post-GDPR Announcement				0.0001	-0.0013***	-0.0007***
				(0.0006)	(0.0004)	(0.0003)
EU × Post-GDPR Effective	-0.0129***	-0.0014	-0.0007			
	(0.0039)	(0.0025)	(0.0017)			
EU × Post-GDPR Announcement				-0.0143***	-0.0026	0.0003
				(0.0041)	(0.0025)	(0.0017)
DI Score × Post-GDPR Effective	0.0052***	0.0051***	0.0029***			
	(0.0004)	(0.0003)	(0.0002)			
DI Score × Post-GDPR Announcement				0.0061***	0.0040***	0.0021***
				(0.0004)	(0.0003)	(0.0002)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,602,696	1,602,696	1,602,696	1,305,942	1,305,942	1,305,942
Adjusted R^2	0.649	0.622	0.597	0.595	0.549	0.509

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 1: Share of Employees in Compliance Roles by Year

This figure presents the share of employment spells classified as being compliance-related in the first 300 employment spells of any given firm, sorted by the year in which the position started. The top graph denotes the share for data-intensive startups, the bottom graph for non-data-intensive startups (as classified by Claude 4.5 Sonnet). The vertical lines indicate 2016, when GDPR was announced and 2018, when GDPR was effective. Note: These shares only cover employment spells from workers included in Revelio Labs. The shares are not directly comparable. The sample period is 2010-2025.

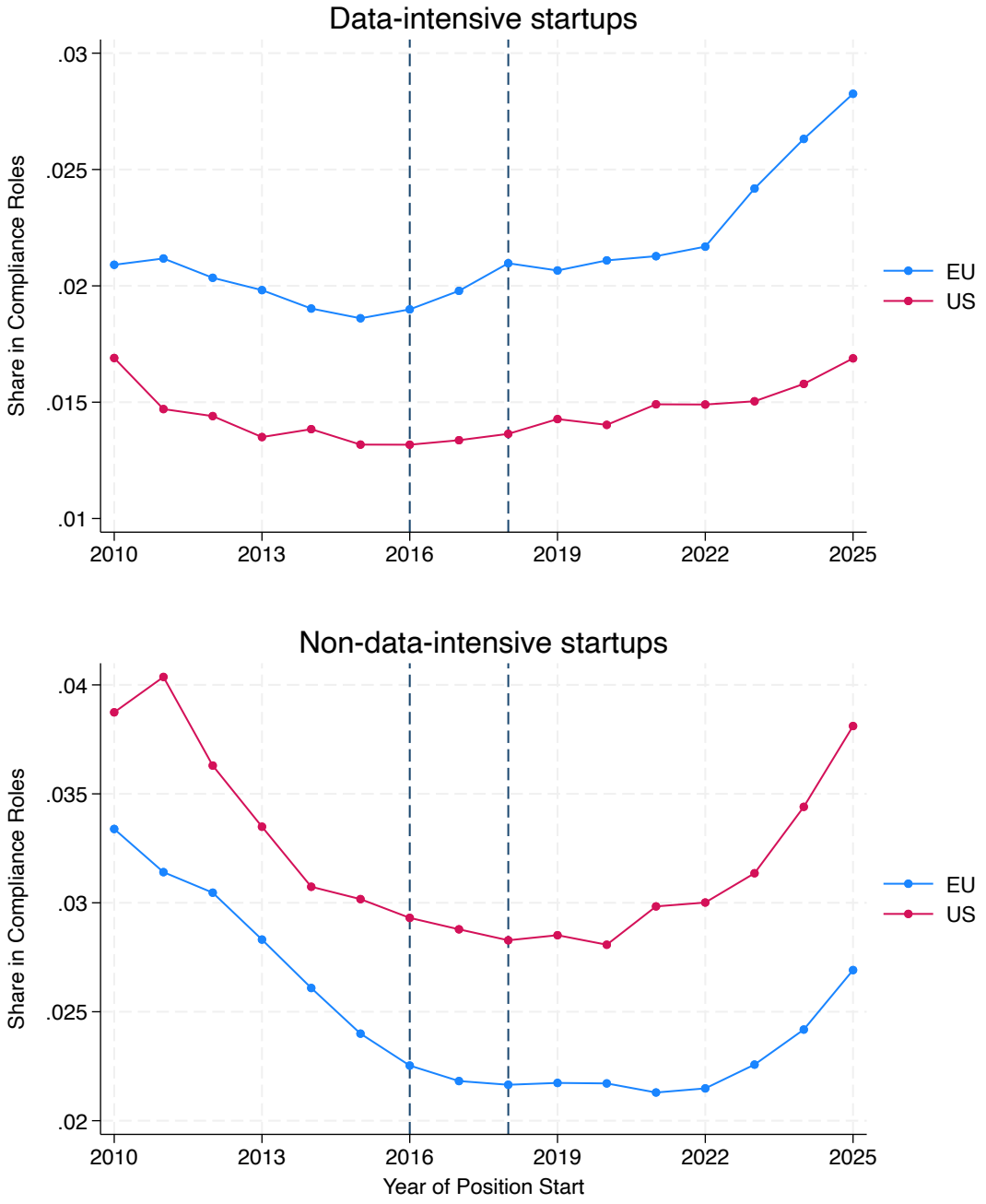


Figure 2: Share of Employees in Compliance Roles - Dynamic DiD / Placebo Treatment

Top Graph: Dynamic DiD The graph shows coefficients and 95% confidence intervals from a regression of the share of employment spells in compliance on $EU \times Year$ interactions, with firm fixed effects. The year 2015 is the omitted category, and all coefficients are plotted relative to 2015. The green vertical line marks GDPR announcement (2016) and the red line marks GDPR implementation (2018). Standard errors are clustered by firm.

Bottom Graph: Placebo Treatment Years This panel re-estimates the main specification from Table 3 using different hypothetical treatment years. For each year on the x-axis, I estimate a separate regression using a 7-year window centered on that year, with $Post = 1$ for employment spells starting after the hypothetical treatment date. The plotted coefficients show the $EU \times Post$ interaction term with 95% confidence intervals. The green vertical line marks the GDPR announcement (2016) and the red line marks GDPR implementation (2018).

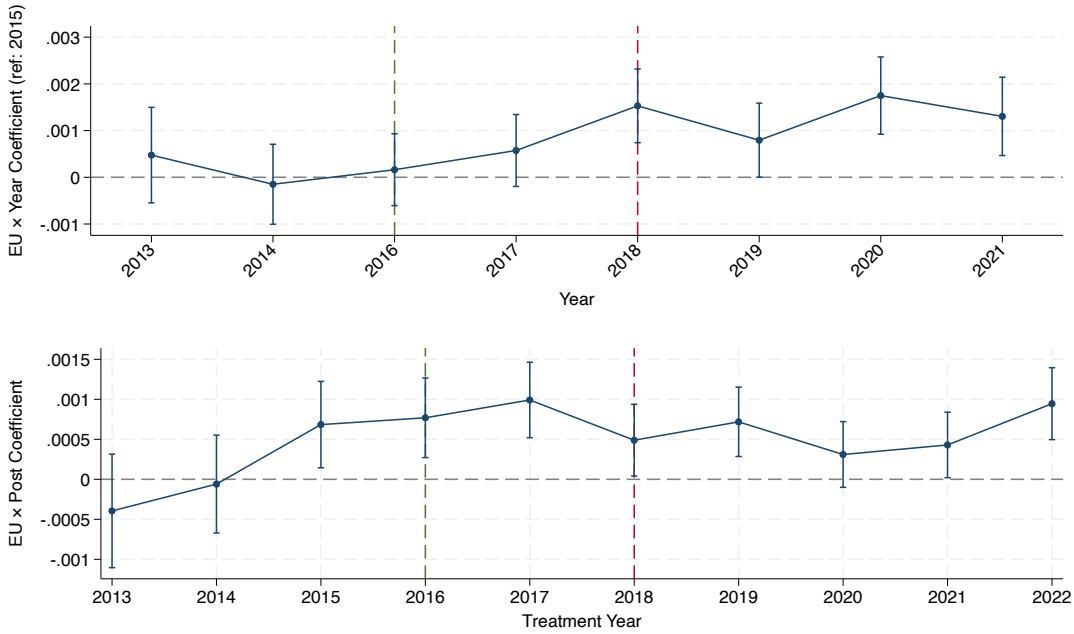


Figure 3: Share of Startups Achieving High Growth by Year

This figure presents graphs showing the share of high-growth data-intensive startups reaching a threshold (50, 150, 300) number of employment spells within four years from the total number of high-growth startups founded in a given year. The vertical lines indicate 2016, when GDPR was announced and 2018, when GDPR was enacted. Note: This graph should not be interpreted as firms being more likely to reach x employees in later years. LinkedIn coverage has increased.

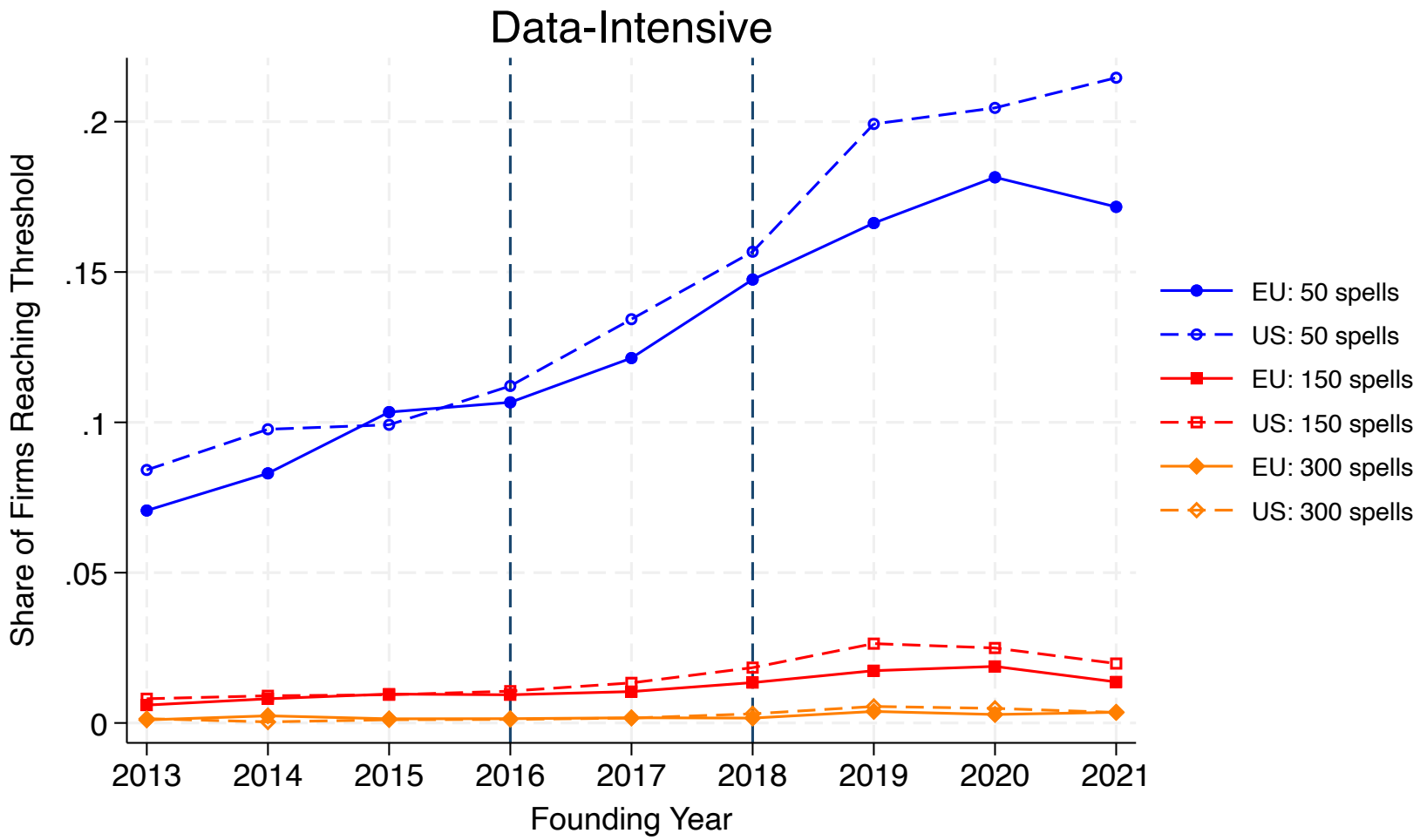


Figure 4: Share of Startups Achieving High Growth by Year - Non-Data-Intensive

This figure presents graphs showing the share of high-growth non-data-intensive startups reaching a threshold (50, 150, 200, 300) number of employment spells within four years from the total number of high-growth startups founded in a given year. The vertical line indicates 2018, when GDPR was enacted.

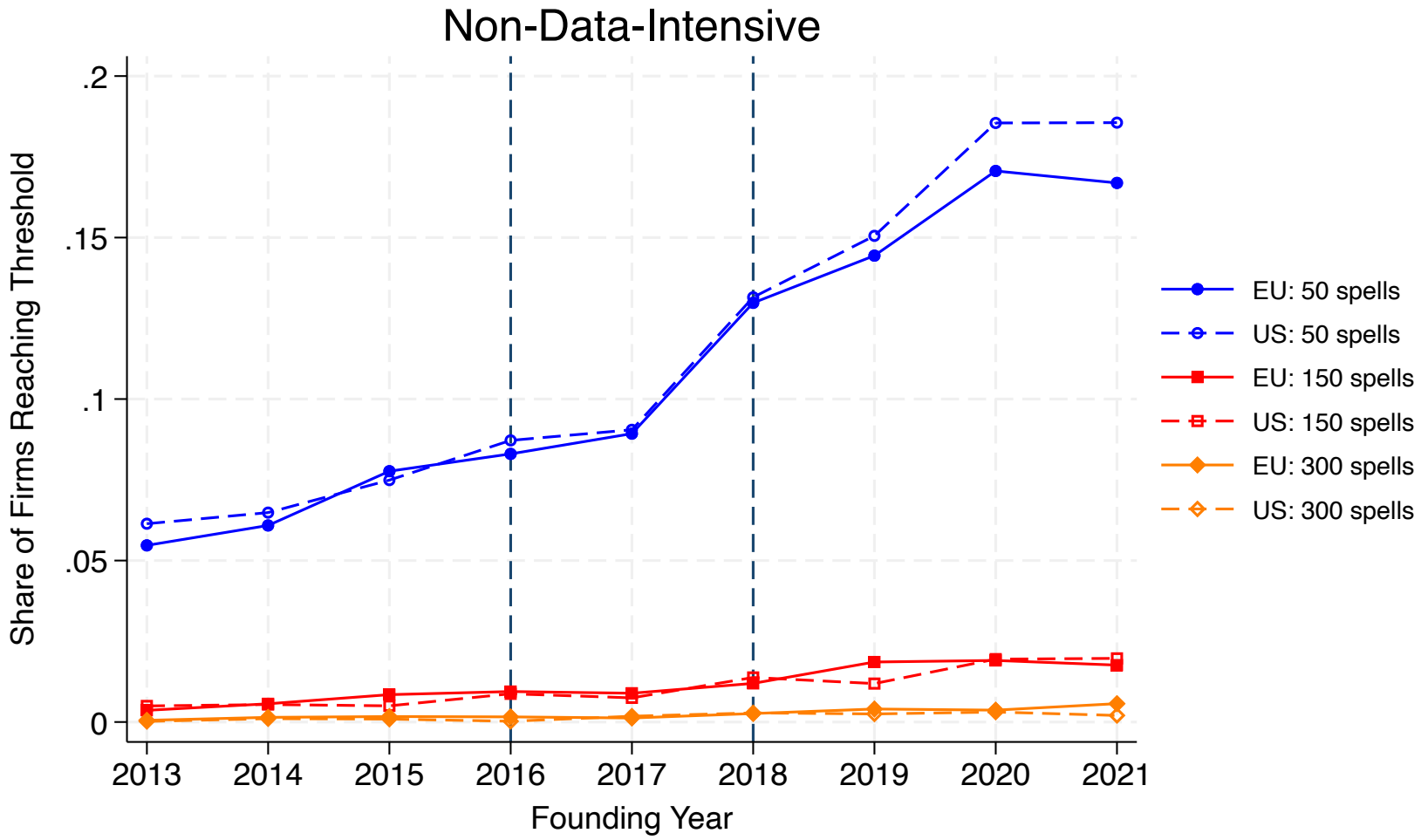
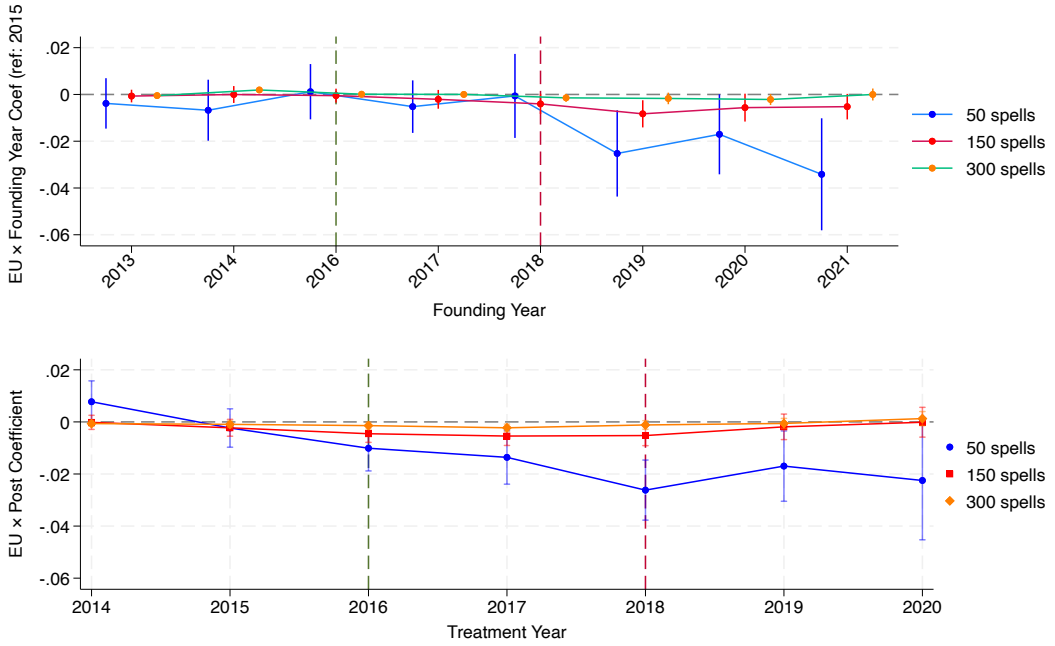


Figure 5: Firm Growth Thresholds - Dynamic DiD / Placebo Treatment

Top Graph: Dynamic DiD The graph shows coefficients and 95% confidence intervals from a regression of the probability of reaching a threshold number of employment spells within four years of founding on *EU* × *Founding Year* interactions, with sector fixed effects. The year 2015 is the omitted category, and all coefficients are plotted relative to 2015. The green vertical line marks GDPR announcement (2016) and the red line marks GDPR implementation (2018). Standard errors are clustered by RICS K400 sector. The three series correspond to different employment-spell thresholds: 50, 150, and 300 employees.

Bottom Graph: Placebo Treatment Years This panel re-estimates the main specification from Table 6 using different hypothetical treatment years. For each year on the x-axis, I estimate a separate regression using a 7-year window centered on that year, with *Post* = 1 for firms founded after the hypothetical treatment date. The plotted coefficients show the *EU* × *Post* interaction term with 95% confidence intervals. The green vertical line marks the GDPR announcement (2016) and the red line marks GDPR implementation (2018). The three series correspond to different employment thresholds: 50, 150, and 300 employment spells.



Appendix

A.1 Data Processing and Filters

This section explains the procedures I use to clean the data. It is intended as a complement to Section 3, and meant to faciliate replication.

A.1.1 Data download

As mentioned in Section 3, I use data from several tables from Revelio. I download the data via WRDS.

I download the user education and company information tables directly in full from WRDS's web interface. However, this is not possible for the individual_positions file. I download this table in chunks (ranging in size from annual to monthly, based on the start date of a job spell) from the WRDS SQL table. The data in this version of the paper were downloaded in December 2025, meaning that the data extend to job spells starting in October 2025.

The main data sources are:

1. **Revelio Labs Individual Positions Files**: Quarterly/monthly files containing individual employment records (2010–2025)
2. **Revelio Labs Company File**: Firm-level information including founding year, headquarters location, and industry classification
3. **Revelio Labs Education File**: Individual education records with degree fields and institutions
4. **O*NET Version 30**: Occupational task data used to construct regulation-intensity indices
5. **AI-Generated Role Classifications**: Claude LLM classifications of job titles into compliance/core operations categories

A.1.2 Industry Data Intensity Classification

I classify all of Revelio's RICS k400 industries into data-intensive and non-data-intensive. The list of industries is obtained from the Company File. The prompt used is in Appendix A.2. I create the `rics_k400_data_intensive.csv` file.

A.1.3 Company Filters

The goal of this section is to identify high-growth startups in certain geographic areas. I do this filter in the `company_file`.

I start with a geographic filter.

A.1.3.1 Geographic Filtering

I filter based on the `hq_region` variable in Revelio's company file. I start by including the regions below and then excluding specific countries. I drop any company that does not meet the geographic criteria below.

Included Regions Western Europe, Southern Europe, Northern Europe, Northern America, Eastern Europe

Excluded Countries Albania, Andorra, Belarus, Bermuda, Bosnia and Herzegovina, Kosovo, North Macedonia, Liechtenstein, Moldova, Montenegro, San Marino, Serbia, Switzerland, Ukraine, Russia, Vatican City

Special Cases Norway, Iceland, and United Kingdom are retained due to GDPR compliance or equivalence agreements.

A.1.3.2 High-Growth Data-Intensive (HGDI) Startup Identification Criteria

Next, I attempt to eliminate non-high-growth startups. As I do not have survey data on growth intentions, I use *ex post* firm outcomes.

In order to do this, I need to use the `individual_positions_(year)` files. As these are chunked, I open each one sequentially and aggregate the data at the firm-level. For

each firm-year combination, I calculate the total number of new employment spells at the company level (`total_starts`) and merge in company information. In this merge, I drop all companies not headquartered in the countries listed above.

This creates files called company_(year) where (year) ranges from 2010 to 2025. I aggregate these into one file called `company_all_years`.

I then classify firms as startups.

A firm is classified as a startup if it satisfies **all** of the following conditions:

1. **Founding Year**: Founded 2010 or later (Revelio data)
2. **First-Year Filter**: Fewer than 100 employment starts in the first year (to exclude acquisitions and mergers)
3. **Sample Duration**: Appears in the sample for at least 3 years
4. **Scale Filter**: Fewer than 1,000 employment starts in any single year (excludes gig-economy/platform firms)
5. **Growth Requirement**: Maximum annual starts exceeds first-year starts by more than 5 employees
6. **Name Filter**: Company name does not contain "Stealth", "stealth" or equal "None (7648393)"

I drop all firm-years which do not meet all of the above conditions.

A.1.3.3 Data Intensity Classification

I then merge in the data intensivity scores from the `rics_k400_data_intensive.csv` file.

The following variables are brought in:

1. Initial classification performed by Claude LLM based on industry descriptions
2. Binary indicator created: `rics_k400_data_intensive` $\in$ {Yes, No}
3. Continuous score: `di_score` (0–10 scale)
4. Justification: One-sentence text explaining the classification

A.1.3.4 Manually Excluded Industries

Finally, all firm-years in the following industries are dropped regardless of AI classification:

- Culture and Entertainment
- Humanitarian Services
- Education Services
- Public Sector Management
- Professional/Trade Associations
- Miscellaneous
- Hospitality and Tourism Management
- Commercial Aviation

A.1.3.5 Variables Created

Variable	Description
`n_starts_year`	Number of employment starts per firm-year
`first_year`	First year firm appears in data (note: this variable will be furthered refined later for the analysis file)
`n_starts_first_year`	Employment starts in firm's first year (same note as above)
`n_years_in_sample`	Number of years firm appears in sample
`n_starts_max`	Maximum annual starts across all years
`n_starts_diff`	Growth indicator: max starts minus first-year starts
`di_score`	Data intensity score (0–10)
`di_category`	Data intensity category

A.1.3.6 Output Files

- `startup_di_all_years.dta`: Data-intensive startups (primary sample)

- `Placebo/startup_non_di_all_years.dta`: Non-data-intensive startups (placebo sample)

A.1.4 Job Spells Data

I start by loading the Revelio position-level files (`individual_positions_(date)`). There are many of these, the data are quarterly/monthly. Each file is roughly 3-8gb.

I load the data and keep the variables listed below.

A.1.4.1 Variables Retained

- `rcid`: Revelio company identifier
- `user_id`: Individual identifier
- `position_id`: Position identifier
- `startdate`, `enddate`: Employment dates
- `role_k1500_v2`, `role_k17000_v3`: Role classifications
- `onet_code`: O*NET occupation code
- `region`: Employee region
- All company metadata from Step 1

I merge this file based on the Revelio Company ID with the `startup_di_all_years.dta` and drop all employment spells not associated with a data-intensive firm. In a separate file for the "placebo" (non-DI) analysis, I match with the non-data-intensive startups file `Placebo/startup_non_di_all_years.dta` and only keep workers working for non-DI startups.

A.1.4.2 Data Cleaning: Deleting "Too Early" Joiners

Employment spells are numbered sequentially within each firm by start date:

```
sort rcid startdate
```

```
bys rcid (startdate): gen emp_num = _n
```

There are many firms where the first employee joins way before any other employee and is in a different region. These are likely to be data errors, or users incorrectly entering their employment history into LinkedIn.

First employees (`emp_num` = 1) are dropped if:

1. Start date is more than 500 days before the second employee, **AND**
2. Employee is from a different region than the second employee

This addresses LinkedIn data entry errors where incorrect early employment records appear.

A.1.4.3 Output Files

- `individual_positions_di_startups.dta`: All individual positions at DI startups (and separate non-DI placebo file)

A.1.5 Merge Role Classifications (K17000, K1500 and O*NET)

A.1.5.1 Classification scripts

I use the Claude API to classify each unique K17000 and K1500 job classification into one of three categories.

A.1.5.2 Role Classification Categories

Category	Description
CORE_OPERATIONS	Core business functions essential to firm operations
COMPLIANCE	Regulatory, legal, data protection, and compliance roles
OTHER	Administrative, support, and unclassified roles

A.1.5.3 Variables Created

Variable	Description
is_compliance_k17000	Binary: 1 if compliance role
is_core_k17000	Binary: 1 if core operations role
is_other_k17000	Binary: 1 if other role
cum_compliance_k17000	Cumulative compliance hires by firm prior to current emp_num
cum_core_k17000	Cumulative core operations hires by firm prior to current emp_num

A.1.5.4 Cumulative Variable Construction

```
bys rcid (emp_num): gen cum_is_compliance_k17000 = sum(is_compliance_k17000)
bys rcid (emp_num): gen cum_is_core_k17000 = sum(is_core_k17000)
```

A.1.5.5 Merging

I then merge these data using the role codes into the individual_positions_di_startups-file

A.1.5.6 O*NET merging process

I download the O*NET v30 tasks file from the O*NET website. The file contains about 19000 job titles as well as the tasks associated with these titles, and for some tasks, an indicator for whether the task is core or not and the source of the task (expert surveys, surveys of workers etc.). I use Python and the Claude API to replicate the methodology of Trebbi et al. (2026) with some differences to account for differences in the O*NET data between versions. The process is described in Section 3.

I create the following variables:

Variable	Description
RTI	Regulation-Intensity Index: proportion of regulation-related tasks
Total_Tasks	Total number of tasks in occupation
Regulation_Tasks	Number of regulation-related tasks

A.1.5.7 Merging

I then merge these data using the O*NET codes as the matching key into the individual_positions_di_sta file

I create firm-level measures of the cumulative reg_index prior to each new employment spell

```
bys rcid (emp_num): gen cum_regindex_total = sum(RTI)
gen cum_regindex = cum_regindex_total / emp_num
```

A.1.5.8 Output Files

- Updated: individual_positions_di_startups.dta

A.1.6 Identify Compliance Education Backgrounds

I then merge in the educational backgrounds of each worker in my data.

A.1.6.1 Sample Restrictions

- Drop records with missing university information
- Exclude "High School" degree level

A.1.6.2 Educational Background Categories

I create six fields of education from the categories available in Revelio. The created variables and the associated fields are provided in the table below.

Variable	Fields Included
`isbusiness`	Business, Finance, Accounting, Marketing
`islaw`	Law
`ismedicine`	Medicine, Nursing
`isscience`	Statistics, Physics, Mathematics, Chemistry, Biology, Architecture
`isengineer`	Engineering
`isit`	Information Technology

Workers may have education in multiple fields. Those whose education history does not provide the field of education are left uncoded.

A.1.6.3 Merge Strategy

- Deduplicate to one row per `user_id`
- Merge into `individual_positions_di_startups.dta` on `user_id`
- Keep: master and matched observations

A.1.6.4 Output Files

- Updated: `individual_positions_di_startups.dta`

A.1.7 Define Analysis Variables

A.1.7.1 Geographic Classification

I create a dummy variable that takes the value of 1 if a firm is headquartered in the EU, UK or Norway. I also create a similar dummy for the US, and third countries (not included in the analysis)

A.1.7.2 EU Indicator

```
gen is_eu = 1 if inlist(hq_region, "Western Europe", "Southern Europe",
                        "Northern Europe", "Eastern Europe")
replace is_eu = 0 if inlist(hq_country, "Switzerland", "Monaco")
```

A.1.7.3 US Indicator

```
gen is_us = (hq_country == "United States")
```

A.1.7.4 Headquarters Group

```
gen hq_group = 1 if is_eu == 1      // EU
replace hq_group = 2 if is_us == 1  // US
replace hq_group = 3 if missing(hq_group)  // Other
```

A.1.7.5 Founding Year Variables

I identify the start date of the firm by the date of the start of the first employment spell. I create a variable `year_founded_data` to incidate this.

A.1.7.6 GDPR Timing Variables

I define the implementation of GDPR in two ways: based on the implementation and announcement dates. For most of my analyses, I use the implementation date.

Two key dates define GDPR timing:

- **Announcement**: April 27, 2016 (regulation adopted)
- **Effective**: May 25, 2018 (enforcement begins)

As most of my analyses are based on annual data, I define treatment based on whether a firm is founded after GDPR implementation (2019 or later) or whether an employee joins after GDPR implementation.

Variable	Definition	Description
`isgdpr_effective_data`	1 if `year_founded_data` > 2018	Firm founded after GDPR is in effect (2019 or later)
`isgdpr_effective_data`	1 if `year_founded_data` > 2018	Firm founded after GDPR is in effect (2019 or later)
`isgdpr_effective_startdate`	1 if employment spell start year > 2018	Employment spell starts after GDPR is in effect (2019 or later)

A.1.7.7 Output Files

- **Final analysis dataset**: `analysis_file_individual_positions_di_startups.dta`

A.1.8 Summary of Data Pipeline

Step	Input	Output	Key Filter
1	company_file.csv, positions files	startup_di_all_years.dta	DI industry, 3+ years, <1000 starts
2	positions files, startup list	individual_positions_di_startups.dta	DI startups only
3a/3b	Role classification CSVs	(updates existing)	K1500 & K17000 categories
3c	O*NET V30	(updates existing)	RTI by occupation
4	education_file.csv	(updates existing)	Education fields

Step	Input	Output	Key Filter
5	All previous	analysis_file_*.dta	Final analysis variables

A.1.9 Final Sample Characteristics

A.1.9.1 Unit of Analysis

Individual employment positions at data-intensive startups

A.1.9.2 Time Period

2010–2025 (employment start dates). All analyses in the paper study the period 3 years either way of GDPR (2015-2021), either based on firm founding or employment-spell start.

Geographic Coverage

- EU: Western, Southern, Northern, and Eastern Europe (excluding Switzerland, Monaco)
- US: United States
- Other: Remaining countries in sample

A.2 Prompts used to classify industries and roles

This section presents the prompts used to classify roles and industries.

A.2.1 Data-intensive industries

The following prompt was used with Claude 4.5 Sonnet on Monday, January 4th, 2026:

```
"""You are classifying industries based on their reliance on
   personal data for core business operations.

Rate each industry on a scale of 0-10, where:
```

- 10 = Extremely data-intensive (core business model requires collecting, processing, or analyzing large amounts of personal data)
- 5-7 = Moderately data-intensive (uses personal data but not central to business model)
- 0-4 = Low data-intensity (minimal reliance on personal data)

Consider factors like:
- Does the industry collect, store, or process large amounts of personal data?
- Is personal data central to the value proposition?
- Does the business model depend on analyzing user behavior or personal information?

For each industry, provide:
1. data_intensive: "Yes" or "No"
2. score: Integer from 0-10 (0=not data-intensive, 10=extremely data-intensive)
3. category: One of: "High", "Medium", "Low", or "Not Data-Intensive"
4. reasoning: Brief 1-sentence explanation

Respond ONLY with valid JSON in this exact format:

```
{
  "data_intensive": "<Yes or No>",
  "score": <integer 0-10>,
  "category": "<category>",
```

```
  "reasoning": "<brief explanation>"
}"""
```

A.2.2 Compliance-related roles

The following prompt was used to classify roles into compliance-related roles and others. It was run with Claude 4.5 Opus on the 23rd of December, 2025 on both the Revelio k1500 role categories and the k17000 role categories

```
    """You are analyzing job roles for a research study on
       workforce composition in data-intensive startups,
       particularly around GDPR compliance (May 2018).

Please classify each of the following {len(roles)} job roles into
   ONE of these three categories:

**CORE_OPERATIONS**: Roles directly involved in creating,
   delivering, or selling the company's products/services
- Examples: Software Engineer, Data Scientist, Product Manager,
   Sales Representative, Marketing Manager, Designer, Customer
   Success, Operations Manager

**COMPLIANCE**: Roles focused on legal compliance, data protection
   , privacy, HR policy, or regulatory matters
- Examples: Privacy Officer, Data Protection Officer (DPO),
   Compliance Manager, HR Compliance, Legal Counsel, Risk Manager,
    Ethics Officer, Regulatory Affairs
```

OTHER: All other roles including general administration, finance, facilities, etc.
- Examples: Accountant, Financial Analyst, Office Manager, Facilities Manager, Executive Assistant, Finance Director

IMPORTANT CLASSIFICATION RULES:

1. Regular HR roles (HR Manager, Recruiter, HR Business Partner) should be classified as OTHER, not COMPLIANCE
2. Only classify HR as COMPLIANCE if the role explicitly mentions compliance, policy, legal, or regulatory duties
3. Data Scientists, Data Engineers, Data Analysts are CORE_OPERATIONS (they build products, not compliance)
4. Only classify as COMPLIANCE if the role name explicitly indicates legal/regulatory/privacy/compliance focus
5. When in doubt between COMPLIANCE and OTHER, choose OTHER (we want a conservative compliance definition)

Please respond with a valid JSON array where each object has:
- "role": the exact role name from the input
- "category": one of "CORE_OPERATIONS", "COMPLIANCE", or "OTHER"
- "reasoning": brief 1-sentence explanation (max 15 words)

Here are the roles to classify:

{json.dumps(roles, indent=2)}

Respond ONLY with the JSON array, no other text."""