Unlocking the Data Advantage: M&A Likelihood and Performance of Data Intensive Companies

author(s) *

January 30, 2024

ABSTRACT

This paper examines M&A activity of firms in the data economy. I develop text-based measures for data intensive companies, i.e., firms that are rich in data resources and capabilities, and relate them to M&A likelihood and performance. Results indicate that data intensity relates to a higher likelihood to become acquirer or target. The largest share of increased acquirer likelihood can be attributed to conglomerate deals. Targets of data intensive acquirers are more often small, foreign, and non-public. While investors do not expect superior performance at deal announcement, data intensive acquirers demonstrate superior long-term stock returns in the years after deal announcement. This study contributes to existing literature by proposing new measures for identifying data intensive companies and investigating these firms' M&A activity and performance.

JEL Classification: G34, L25, O30

Keywords: Data Economy, Mergers & Acquisitions, Natural Language Processing, Deal Likelihood, Deal Performance

* Data firm measures and Python code will be accessible online upon finalization of the project.

INTRODUCTION

"[D]ata are a natural resource, much like oil, which can be owned and traded [...]. Businesses are [...] facing a digital reversal. Many firms want to use data to infuse their corporate applications with AI. They have built central repositories such as 'data lakes', which hold all kinds of digital information. Such systems are of limited use, however, if a firm and its employees lack the required skills, refuse to believe the data or even to share them internally." – The Economist (2020)

The growth of the data economy has sparked increased interest in the role of data, as well as the skills and tools required to analyze it, in business strategy. Data and data transformation capabilities can create a competitive advantage (e.g., McAfee et al. 2012). When striving to obtain, expand, or keep such a competitive advantage, data resources and capabilities can be appealing transaction assets that can motivate investments in mergers and acquisitions (M&A).

For instance, anecdotal evidence suggests that big tech firms have been investing heavily in M&A. According to Thomson Reuters Eikon, the 'big five' U.S. tech giants *Alphabet/Google*, *Apple*, *Meta/Facebook*, *Amazon*, and *Microsoft* completed 553[1] deals in 2006-2022. Out of these deals, only 95 report a deal value, which collectively amounts to $266 billion. Most acquisition targets (95%) were private or subsidiary firms, suggesting that they were rather small when acquired. Almost 40 percent of these five firms' acquisitions were outside of their primary industries (measured on a 3-digit SIC code-level), indicating diversification (i.e., conglomeration) or vertical integration. Although looking at the 'big five' can provide striking examples and deal values, this leaves out a large portion of other big tech companies that are also investing heavily in M&A.

In line with these discussions, existing research on data intensive firms is mostly case-based or anecdotal, investigating M&A activity of specific, well-known companies (e.g., Argentesi et al.

[1] This number includes all completed transactions by these five firms in the database, independent of the acquired share or other filter criteria applied in the study later on.

2021; Gautier and Lamesch 2021; Prado and Bauer 2022). Other studies are theoretical, modeling the effects of big tech acquisitions on welfare, innovation, and competition (e.g., Cabral 2023; Katz 2021; Motta and Peitz 2021). More research covers the role of technology in acquisitions (e.g., Ahuja and Katila 2001; Sears and Hoetker 2014), measuring technological stock, capabilities, and acquirer-target overlap based on patents and R&D expenditures (e.g., Kaul 2012; Sears and Hoetker 2014). However, resources and capabilities in the digital sector also include data, as well as analytical skills, tools, and equipment (Bourreau and De Streel 2019), which cannot be fully captured by the technology measures used in prior research.

Despite these streams of literature and discussions about market power and acquisition activity by tech giants, we know relatively little about whether these phenomena only hold for few, well-studied firms and whether deals by data intensive firms are valuable investments. As data and data-related capabilities become increasingly important for corporate takeovers, this study adds empirical evidence on the relation between a firm's data intensity and its M&A activities.

I apply textual analysis to measure to what extent a firm's business relies on generating data, using and/or providing skills, tools, and other capabilities that help transforming data into valuable insights. I analyze 10-K business descriptions and risk factors (Items 1 and 1A) for U.S. listed companies for the fiscal years 2006-2021, building on previous studies that use 10-K business descriptions (Item 1) to derive text-based measures (e.g., Hoberg and Phillips 2010, 2016).

The developed measures for data intensity reflect expected characteristics of data firms, such as their frequent occurrence in the Computer Software and Hardware industries. Some of the highest scoring companies are *Intuit*, *Cisco*, and *F5*. Highly data intensive firms include companies active in consumer- and/or business oriented markets. Hence, top-scoring firms do not necessarily collect vast amounts of consumer data themselves, but are particularly strong in data capabili-

ties, such as in cloud services, data analytics, or cyber security. The proposed measures thus not only capture data ownership, but also corresponding capabilities.

Using these measures in multivariate analyses, I show that higher data intensity is associated with a higher likelihood of becoming an acquirer or target. This indicates that data intensive firms are both active acquirers and attractive acquisition targets. Data intensive acquirers particularly often invest in targets that are small, non-listed, foreign, young, and outside the acquirer's core business (conglomerate deals). At the same time, vertical deals are less frequent for data intensive acquirers compared to horizontal or conglomerate transactions. The results suggest that, while market power-driven incentives and pre-emptive acquisition activity cannot explain the increased acquisition activity (though they still might be the driver behind individual acquisitions), conglomerate incentives seem to be a more important motivation. Data intensive acquirers do not experience higher announcement returns, but a superior long-term stock performance.

This study helps to understand not only whether data intensive firms are more active in acquisitions and whether they experience superior M&A performance, but also which motives likely explain increased M&A activity. It contributes to existing literature in three ways. Firstly, while prior studies on data firms or assets use small samples based on surveys (e.g., Brynjolfsson et al. 2011), technical skills of employees using LinkedIn data (e.g., Tambe 2014), or job postings in a single industry (e.g., Abis and Veldkamp 2023); this study proposes new measures for identifying the extent to which firms possess data resources and capabilities, which provides an opportunity for further empirical investigations of data intensive businesses. Secondly, it investigates M&A activity by data intensive firms for a large number of companies. It complements existing studies which are mostly case-based or provide anecdotal insights (e.g., Argentesi et al. 2021; Gautier and Lamesch 2021). The proposed measure provides the opportunity to present large-scale

empirical evidence for the regained importance of conglomeration for data intensive companies. Lastly, it explores how acquisition performance relates to a company's data intensity, an area in which prior literature has focused more on other technological resources and capabilities such as R&D and patents (e.g., Ahuja and Katila 2001; King et al. 2008; Sears and Hoetker 2014).

The study is interesting for both policy makers and managers. The M&A trend in tech firms appears to extend beyond the 'big five', *Alphabet/Google*, *Apple*, *Meta/Facebook*, *Amazon*, and *Microsoft*. At the same time, the frequently discussed motive of pre-emptive buyouts cannot explain the increased M&A activity among data-intensive firms. Rather than focusing solely on acquiring future competitors within their core market, data intensive firms often engage in conglomerate transactions, expanding into (seemingly) unrelated industries. Leveraging their data resources and capabilities, they aim to exploit economies of scope, potentially capitalizing on their market power in new domains. Since potential exploitation of market power when entering new industries typically faces less scrutiny from competition authorities, this study underlines the necessity for revisions in the evaluation of data intensive M&A deals.

Data intensive firms seem to benefit from investing in potentially risky acquisitions of often small, young, and innovative firms. Despite insignificant differences at deal announcement, data intensive firms' long-term stock performance post-acquisition surpasses the performance of less data intensive counterparts. Since data intensity can, for example, improve target selection or the exploitation of M&A benefits post-acquisition, managers of firms low in data intensity should invest in improving data stock and capabilities to catch up with their more data intensive peers.

The next section provides a brief overview on existing literature explaining motives for potentially increased M&A activity by data intensive firms and resulting performance implications. Afterwards, the sample, the measurement of data intensity and other variables, and methods are

introduced. Then, I present descriptive and multivariate analyses of deal likelihood and performance as well as deal, target, and acquirer characteristics. The final section concludes the paper.

LITERATURE REVIEW

Data Intensive Firms

The starting point to discuss implications of a firm's data intensity, is to define data intensive firms. I build on the intangible asset literature, where data –while being in principle nonrival– is often accumulated and exploited exclusively by companies. The exclusivity does not only stem from the preference of companies to keep them as a business secret, but sharing in particular consumer data is limited due to privacy regulation (Corrado et al. 2022). Hence, data intensive firms are often generators of consumer-related data, which can become a valuable asset when analyzed and incorporated in business activity (e.g., Argentesi et al. 2021; Lambrecht and Tucker 2017). This can mean, for example, better targeting advertisement or product recommendations to consumers or refining the estimation of demand forecasts to among others set prices and improve warehouse efficiency. For this transformation data needs to be stored, managed, secured, analyzed and/or fed into algorithms with the goal to use the resulting insights in future business activity. A vast amount of firms active in the data economy might not themselves generate a lot of data, but provide the data capabilities needed for this transformation process. The following section discusses how such resources and capabilities can be valuable assets in an M&A context.

Deal Likelihood

In general, acquisition motives typically discussed evolve around value creation through cost-reducing or revenue enhancing synergies and managerial incentives (e.g., Trautwein 1990). While the former is theorized as beneficial for the acquisition parties, the latter is often perceived as value destroying. Value creation can occur among others through exploiting economies of scale

and scope, acquiring resources or capabilities that are difficult or slow to produce internally (e.g., by acquiring vertically related firms, firms that employ talent, or firms that own valuable resources), enhancing market shares, or efficiency improvement and tax savings.

I argue that some of these M&A motives are particularly pronounced for data intensive companies, resulting in increased incentives to invest in M&A. These resolve around the following three aspects. First, if firms want to improve their data resources or capabilities and cannot/do not want to invest the effort and/or time to develop them internally, they can retrieve them by acquiring data intensive targets. Acquirers in such transactions can be more or less data intensive themselves, attempting to expand existing or obtain new data resources and/or capabilities. Second, the market characteristics in which data generating firms operate are often accompanied by increased incentives for rapid growth, among others through M&A. Such deals could be observed in an increase in horizontal acquisitions. If targets are particularly young and innovative, such deals could even represent pre-emptive buyouts. Third, data generating firms can use their data insights to find the most attractive industries to enter, among others through conglomerate deals. Also firms strong in data capabilities, such as data analytics, could use such deals to exploit economies of scope to new data sets in (seemingly) unrelated industries.

To provide a more detailed explanation of these reasons, I begin by addressing the potential asset value of data resources and capabilities. Acquisition of such valuable assets through M&A is one of the main motives for pursuing transactions, especially for technology firms (e.g., Graebner et al. 2010). Data are valuable information, representing intangible assets that results from a firm's business activity (e.g., Farboodi et al. 2019; Veldkamp 2023). Such data resources typically require competencies to contribute to a firm's value generation (e.g., Amit and Schoemaker 1993; Barney 1991; Lambrecht and Tucker 2017). Applying analytical skills and tools, companies use

data to improve their offers and internal processes to generate income and gain a competitive advantage (e.g., Argentesi et al. 2021). Acquisitions can help to obtain these capabilities, especially if they are difficult to develop internally (e.g., Eisenhardt and Martin 2000). Hence, not only increasing the amount and scope of data but also adding or expanding analytical skills and tools, can thus be an acquisition motive for data-driven firms. Even businesses that to date would not define themselves as data-driven, might acquire data-capabilities to enable exploiting potentially unused data resources and thus become more data intensive in the future.

Furthermore, many data generating businesses have a strong interest to increase market power due to the particular nature of markets, in which data intensive firms often operate (e.g., Bourreau and De Streel 2019). Their market characteristics can increase incentives to grow, among others by M&A. For data generating firms, building a large data stock is typically expensive in the beginning, but gets cheaper with increasing size, up to marginal costs close to zero. Such scale economies are often accompanied by network effects, as many data generating firms' business models are based on platforms, i.e., two-/multi-sided markets, acting as an intermediary between at least two sides of customers (e.g., Campbell et al. 2015). In two-sided markets, data can be, for example, used to better match the different sides through personalized advertisement or product recommendations, or to optimize demand forecasts to set prices and enhance warehouse efficiency. Scale economies combined with network externalities can increase entry barriers and simplify monopolization (e.g., Campbell et al. 2015). Some markets even have 'winner-take-all' characteristics, suggesting that natural monopolies are a probable outcome of competition (e.g., Correia-da Silva et al. 2019; Katz and Shapiro 1985). Especially in case of technology-oriented businesses, acquirers often target young, innovative firms (e.g., Graebner et al. 2010). If these are market power driven, they can even represent pre-emptive buyouts, i.e., buying potential future competitors – sometimes called 'killer acquisitions' (see Cunningham et al. 2021).

Finally, especially in big tech, conglomeration re-gained relevance (e.g., Gautier and Lamesch 2021). Economies of scope are an important driver, in particular when entering new markets. Not only (parts of) analytical know-how and equipment can be easily adapted to expand to related or new digital markets, but also data as a shareable input factor can be used in other areas at basically no further cost (e.g., Bourreau and De Streel 2019). Data can reveal consumer preferences, trends, and through this potentially attractive markets to enter. This information advantage can reduce uncertainties, which makes market entry, directly or through M&A, less risky. Zhu and Liu (2018) report such data-driven advantages for Amazon, which has a higher probability to enter product categories that are popular on its marketplace. The example of Amazon depicts furthermore that some seemingly unrelated big tech acquisitions when comparing commonly used industry classifications might be in fact related in terms of their data scope and capabilities (Bourreau and De Streel 2019). The finding that resource or technological similarity increases acquisition likelihood and performance (e.g., Bena and Li 2014; Chondrakis 2016; Hoberg and Phillips 2010; Wang and Zajac 2007) can thus hold even when building digital conglomerates by expanding to seemingly unrelated product markets. Research on the largest tech firms (e.g., Argentesi et al. 2021; Latham et al. 2020) shows that they exhibit a conglomeration motive through their acquisition behavior, often targeting companies outside their core business domains.

These strategic motives and characteristics of data-driven business models overall point to increased acquisition incentives. On the one hand, data intensive firms could be more likely to become acquisition targets since acquirers have incentives to enhance or acquire data resources and/or capabilities to expand their market share or to obtain these assets when internal development would raise difficulties. Additionally, data intensive targets could search for acquirers with better access to customers while boosting the acquirers' data resources or capabilities to improve their offers or enter new markets. On the other hand, data intensive firms could be more likely

to become an acquirer to increase market power, to apply a pre-emptive acquisition strategy, or to further improve their data stock and transformation tools. These arguments would point to an increased incentive to acquire (potential) competitors or firms in closely related industries (i.e., more horizontal deals). Furthermore, the likelihood to become an acquirer could increase due to potentially easier exploitation of economies of scope and lower entry barriers to new markets. This would point to an increased incentive to acquire firms in different industries, which can be (seemingly) unrelated to the data intensive firms' core business (i.e., more conglomerate deals).

Deal Performance

Most of prior M&A literature states that, on average, acquirers do not gain in acquisitions, sometimes with negative, sometimes zero announcement effects, while targets appear to gain value (see e.g., Mulherin et al. 2017 for a literature review). Especially firms that do multiple acquisitions (as, for example, Google with on average almost one acquisition per month) do not necessarily change an investors perception of a company, which can lead to non-significant bidder announcement returns (e.g., Cai et al. 2011; Fuller et al. 2002). Results on long-term stock performance post-acquisition are mixed (e.g., Betton et al. 2008; Martynova and Renneboog 2008) and measurement methods especially of earlier studies were often criticized (e.g., Lyon et al. 1999).

Additionally, scholars acknowledge that there is substantial heterogeneity in deal performance, with several moderating variables discussed in existing studies. For example, acquiring technological or digital resources through M&A as well as higher resource complementarity and relative technological proximity appear to positively affect bidder announcement returns and long-term profitability (e.g., Chondrakis 2016; Hoberg and Phillips 2010; King et al. 2008; Kohers and Kohers 2001; Uhlenbruck et al. 2006). Their reasoning for these results typically builds on growth through innovation potential and a higher strategic fit of complementary firm resources and ca-

pabilities. This fit can improve post-merger integration and improved innovation performance. Particularly if deals are targeted towards improving innovation performance, these acquisitions are rather long-term oriented compared to, for example, acquisitions motivated rather by cost-saving potential or financial synergies (e.g., Rabier 2017).

Another reason for the potentially superior performance of deals by data intensive acquirers lies in their ability to select targets, particularly in those cases in which data intensive firms can use their data to identify trends, find the most profitable customer group and the most attractive markets to enter (e.g., Zhu and Liu 2018). This reduced information asymmetry can help to more accurately evaluate targets and the markets in which they operate, and thus potentially improve acquirer performance (e.g., Bergh et al. 2019).

In summary, the expected relation between data intensity and acquisition performance is less straightforward. On the one hand, data intensive acquirers might acquire more often, potentially lowering the visibility or the impact of a single acquisition on overall performance. On the other hand, digital resources and capabilities acquired in a deal might be value enhancing due to the resource complementarity and the potential exploitation of economies of scale and scope.

The performance implications for data intensive targets seems clearer. If data intensive firms are more attractive acquisition targets compared to less data intensive firms as argued in the previous section, competition for such targets can be high, potentially increasing acquisition premiums and thus improving target deal performance. There is theoretical research against this argumentation for the case when targets are start-ups, indicating that the possibility of a killer acquisition by an incumbent technology firm can reduce acquisition prices, and thus deal premiums, for new entrants (Kamepalli et al. 2022). However, since this study primarily focuses on the target

likelihood and performance of publicly traded firms, which are typically not start-ups, their rationale is only marginally relevant in the context of this research.

To analyze the relation between data resources and capabilities and acquisition likelihood and performance, the first step is to develop a proxy for the data intensity of firms. In the following section, I present a measurement approach based on textual analysis.

DATA AND METHODS

Sample

The sample covers U.S. headquartered firms that published 10-K annual reports in the fiscal years 2006-2021. 2006 is the first fiscal year in which reporting risk factors in a separate item (*1A*) is mandatory for listed companies.[2] This *Item 1A* is exploited in the proposed data intensity measure. Furthermore, to be included, firms must be listed at the NYSE, Nasdaq, or AMEX with available return data in CRSP and valid accounting values in Compustat. Finally, there must be a matching 10-K filing published less than 365 days after a firm's fiscal year end date, resulting in 37,760 firm-fiscal year observations.

Independent Variable – Data Firm Measure

Construction

The idea behind constructing the data firm measures builds on the following assumption: companies that have similar business models and face similar risks compared to a group of identified data intensive companies, are also likely to be data intensive. The definition of data intensive firms builds on the intangibles literature. To identify data intensive firms, I collect merger review cases from the Federal Trade Commission (FTC) and the European Commission (EC), in which

[2] Smaller reporting companies are exempted from this requirement and are therefore largely not covered in this study. Smaller reporting companies have either a market value below $75 million or revenues under $50 million and no public float. For the definition of small reporting companies, see sec.gov/corpfin/amendments-smaller-reporting-company-definition.

press releases or decision texts discussed the aggregation of, or access to, consumer data by the transaction parties. The identified case firms serve as a benchmark for measuring data intensity.

Manual search through official press releases and case descriptions from the FTC (ftc.gov) and the EC (eur-lex.eu) throughout the years 2006-2021 for indications of concerns related to privacy, access, or aggregation of sensitive data in merger review cases identified the following cases[3]:

- Google/DoubleClick acquisition 2007 (FTC 2007; European Commission 2008);
- Facebook/WhatsApp acquisition 2014 (FTC 2014; European Commission 2014);
- Sanofi/Google/DMI joint venture 2016 (European Commission 2016a);
- Microsoft/LinkedIn acquisition 2016 (European Commission 2016b);
- Microsoft/Github acquisition 2018 (European Commission 2018a);
- Apple/Shazam acquisition 2018 (European Commission 2018b);
- Google/Fitbit acquisition 2019 (European Commission 2020);
- Microsoft/Nuance acquisition 2021 (European Commission 2021);
- Meta/Kustomer acquisition 2021 (European Commission 2022).

From these case firms, all publicly listed companies can be used for further analyses, i.e., the acquirers Alphabet/Google, Apple, Meta/Facebook, and Microsoft, as well as the acquisition targets Fitbit, LinkedIn, and Nuance.[4]

To measure the similarity of a firm in my sample to each of the seven identified data intensive companies, I analyze their 10-K business descriptions (Item 1) and risk factors (Item 1A). 10-K business descriptions reflect a company's business model, i.e., which products or services they offer including how they generate income from selling them. Risk factors describe all potential impact factors inside and outside the firm that could affect, for example, a company's earnings or litigation risk as well as how a company conducts its business. Business descriptions (10-K Item 1) have been used in several influential studies, such as Hoberg and Phillips (2010, 2016). I further add the section on risk factors (Item 1A), as prominent data intensive companies discuss

[3] Please find a table with the identifying sections in the sources in Online Appendix A.2.

[4] In additional tests, instead of using the merger case firms' 10-K items as benchmark texts, I use 10-K business descriptions and risk factors of the U.S. firms listed in the Nasdaq Yewno Global Artificial Intelligence and Big Data Index (NYGBIG®). Further robustness tests are conducted and discussed after presenting the results.

potential harm to their business procedures and income, including those resulting from stricter data privacy regulations, such as the European General Data Protection Regulation (GDPR). The items 1 and 1A are extracted from the Loughran and McDonald Stage One 10-K files.

I measure the cosine similarity ($Similarity1_{ij,t}$ and $Similarity1A_{ij,t}$) of each of the seven case firms' 10-K items 1 and 1A (firm i in year t, $i \in [1, 7]$) to each of the remaining firms' 10-K items 1 and 1A in the overall sample (firm j in year t). The overall $Similarity_{ij,t}$ to each of the case firm observations is the cosine similarity to Item 1 ($Similarity1_{ij,t}$) plus the cosine similarity to Item 1A ($Similarity1A_{ij,t}$). More details on the input list of words and the detailed method for calculating cosine similarities are listed in Online Appendix A.1. The average yearly similarity to these case firms $Similarity_{j,t} = \frac{1}{N}\sum_{i=1}^{N} Similarity_{ij,t}$ defines the continuous data intensity measure.[5] For each of the seven case firms, the data intensity reflects the mean similarity to all other case firms. The data intensity measure is normalized to values between 0 and 1. In the following section, I show how the described data intensity measure reflects expected data firm characteristics to justify the suitability of the presented measure.

Verification

The data intensity variable has a slightly right-skewed ($skewness = .42$) distribution with most frequent values around .3, a mean of .39, and a median of .37. A low number of observations are in both extremes with a kurtosis of 2.62. The overall standard deviation is .18.[6]

To show the power of the data firm measures, I start with looking at the firms with the highest data intensity levels. Eight of the top 10 firms in fiscal year 2021 are in the Computer Software industry.[7] Some firms offer software or online tools for consumers (e.g., Intuit). Others offer, for

[5] Note that not all of the seven case firms are in the sample every year. Therefore, the number of benchmark firms varies each fiscal year. The case firms are assumed to be data intensive whenever they are in the data set and thus serve as a benchmark throughout their respective sample periods.

[6] Online Appendix A.3 shows a density histogram of the normalized data intensity scores.

[7] Online Appendix A.4 lists the top 10 data firms including their industry and data intensity score.

example, data management and protection (e.g., AvePoint), cloud computing (e.g., ServiceNow and VMware), networking and IT infrastructure (e.g., Cisco and HPE), or application and network security (e.g., F5). Out of the top 10, Cisco, F5, HPE, Intuit, and ServiceNow are listed in the Nasdaq Yewno Global AI and Big Data Index (NYGBIG®).[8] This index includes 55 U.S. firms, among others Alphabet/Google, Amazon, Meta/Facebook, Apple, and Microsoft. All these five tech giants score in the highest data intensity decile. 37 out of the further 50 companies listed in NYGBIG® in my sample score in the highest data intensity decile as well.

Furthermore, Figure 1 presents the mean and standard deviations of the data intensity scores of all firms within Fama French 49 industries that cover more than 500 observations in my sample. Computer Hardware industry has the highest mean with .68. Computer Software firms rank second, with a mean of .64. Note that Computer Hardware firms show a slightly increasing development over time, surpassing Computer Software firms only in more recent years. Electronic Equipment ranks third with a mean of .59. Industries such as Entertainment, Restaurants, Hotels, and Motels, as well as Petroleum and Natural Gas have low data intensity scores. Furthermore, scores substantially vary within industries, with standard deviations between .06 and .17, indicating that industry classifications can only partially explain data intensity.[9]

INSERT FIGURE 1 HERE

Dependent Variables

In this study, M&A likelihood of U.S. publicly listed firms is the main outcome variable. A binary variable turns to one if a firm was an acquirer (target) within the following firm-fiscal year,

[8] The full list of firms including their data intensity percentile are reported in Online Appendix A.5. Note that the index has been launched in 2018 and the constituents listed here are those of September 2023, so after the sample period. Furthermore, some of the indexed firms were not yet public or partly less data intensive during my sample's time period. Note that inclusion in the NYGBIG® builds on textual analysis of firm's patents, which likely cover some part – but not all – of the data intensity measure presented here. See nasdaq.com for more details on the inclusion criteria.

[9] These industry clusters are also reflected in word clouds generated from bag of words of the top and bottom 100 data intensive firms in the fiscal year 2015, see Online Appendix A.6.

i.e., within 365 days after the fiscal year end date at which firm level variables are measured. Deal data is collected from Thomson Reuters Eikon. Deals are included if the target is a public, private, or subsidiary firm; if the deal status is 'Completed'; if the deal form is 'Merger', 'Acquisition', 'Acquisition of Assets', or 'Acquisition of Majority Interest'; and if the acquirer owned less than 50 percent of the target before and at least 50 percent of the target after the deal.

Further dependent variables are explained when used in additional analyses to understand the motives and performance of M&A deals by data intensive firms.

Control Variables

Several control variables are included in the multivariate analyses (see e.g., Becher et al. 2022; Cremers et al. 2009; Kaul 2012 for similar approaches). They are presented in Table 1 below the dependent variables acquirer and target and the main explanatory variable data intensity. I furthermore include Fama French 49 industry times year fixed effects to account for industry-specific and macroeconomic factors.

Methodology

To estimate the relation between data intensity and acquisition likelihood, I run logistic regressions using firm-clustered standard errors. Target likelihood and acquirer likelihood are estimated in separate regressions. In robustness tests, I also apply a multinomial logistic regression with the categories acquirer, target, or no deal in the dependent variable.

RESULTS

Descriptive Statistics

To get an overview of the sample, Table 1 reports summary statistics of the variables included in the regressions. The share of firm-fiscal year observations involved in acquisitions as an acquirer

(target) is 27 percent (5%). Almost half of the observations report R&D expenditures (48%), resulting in a positively skewed distribution of R&D intensity and R&D stock.Looking at firm's income statements, on average they report $533 million in sales (= $exp(6.258)$), a net income of -1 percent relative to total assets (ROA), and a market equity to net income of 14 (PE ratio). Firms' balance sheets report on average 21 percent tangible assets, 6 percent other intangible assets, and 18 percent cash relative to total assets. Firms are included in Compustat for on average 24 years (age) and mean market concentration is 28 percent (HHI). 44 percent of observations are categorized in the mature lifecycle stage, while 8 percent are in the introduction, 30 percent in the growth, 12 percent in the shake-out, and 6 percent in the decline lifecycle stage.

When comparing the seven case firms' characteristics (in total 65 firm-fiscal year observations, not depicted separately) with the remaining sample, mean acquirer, data intensity, R&D stock, Tobin's Q, cash, ROA, and sales are higher for case firms. No R&D, tangible, leverage, and being in the introduction lifecycle stage are lower for case firms. For target, R&D intensity, HHI, age, PE ratio, intangible, and the other four lifecycle stages growth, mature, shake-out, and decline, 95% confidence intervals of case firms and the remaining sample overlap to a larger extent (p-$values > .05$) and are thus not considered to be different.

The highest pairwise correlation[10] for the acquirer dummy is sales (.25), indicating a positive relation between firm size and acquisition activity. Correlations to the target dummy are comparably low, with the most pronounced correlation to acquirer (-.07). The data intensity variable has a high positive correlation to R&D stock (.54), suggesting that data intensive firms are investing more in innovation activity. Accordingly, the highest negative correlation of data intensity is recorded for no R&D (-.46). The correlation between the main explanatory variable data intensity and the dependent variables acquirer (target) is .16 (.04). For the acquirer dummy, data

[10] The entire correlation table is presented in Online Appendix A.7.

intensity is the variable with the second highest absolute correlation after sales. For the target dummy, data intensity ranks fourth in absolute correlation after acquirer, sales (-.05) and age (-.05). These are first indicators for a positive relationship between data intensity and deal likelihood, which is further investigated in multivariate analyses in the next section.

Multivariate Analysis

Acquirer and target likelihood

This section reports the results of the multivariate analyses estimating the relation between acquisition likelihood and data intensity while controlling for other impact factors. Table 2 reports the results, which are presented as marginal effects with standard errors in parentheses.

The results of the data intensity variable show a positive relation to the likelihood to become an acquirer or target (columns (2) and (7), $p\text{-}values = .00$). In economic terms, acquirer (target) likelihood is 2.0 percent (0.6%) higher when data intensity increases by .1. For example, a company at the 75-percentile of the data intensity variable ($p75 = .50$) has a 4.9 percent (1.4%) higher likelihood for becoming an acquirer (target) than a firm at the 25-percentile ($p25 = .25$). These results indicate that data intensive firms are more likely to be acquirers and are more attractive acquisition targets than firms that are less data intensive.

Including the data intensity variable improves goodness of fit when analyzing acquirer likelihood, as shown by the increase of the Pearson chi^2 statistic and pseudo R^2 (column (1) vs. (2)). This indicates that the data intensity variables can explain a proportion of acquisition likelihood beyond industry fixed effects. In the analyses of target likelihood pseudo R^2 increases and Pearson chi^2 statistic decreases when adding the data intensity variable (column (6) vs. (7)).

Higher R&D stock, no R&D, ROA, sales, intangible assets, and growth lifecycle stage show a positive relation to acquirer likelihood. Tangible assets, cash, leverage, age, and shake-out life-

cycle stage are negatively related to acquirer likelihood. Target likelihood increases with higher R&D stock, leverage, and being in a mature or shake-out lifecycle stage. A negative relation to target likelihood can be observed for R&D intensity, Tobin's Q, cash, sales, HHI, and firm age. Hence, younger, smaller companies, and firms with less growth perspective appear to be attractive acquisition targets. Growing, larger, more profitable firms, and firms with high intangible and low tangible assets as well as low debt are more likely to become acquirers.

I conduct several robustness checks for the main regression analysis. In these robustness checks, results for the data intensity measure hold and are economically similar to the ones reported in Table 2. First, instead of separate logistic regressions for acquirer and target likelihood, I run a multinomial logistic regression with a multinomial dependent variable (for the three categories acquirer, target, or no deal). Firms that appear as acquirers and target in the same year are categorized as targets in this variable. In the analyses reported in Table 2, both acquirer and target dummies turn to one for such firms. Second, I exclude a selection of industries, in particular those with special M&A regulation, i.e., where some federal or state authority issues licenses for pursuing a particular business activity. Typically, if licenses change its possessor, the specific emission agency must approve. This is true for financial firms (e.g., in the case of banking the respective authority is the Federal Reserve Board), businesses active in telecommunication (Federal Communications Commission), energy (e.g., Federal Energy Regulatory Commission), and transport (e.g., Surface Transportation Board). Hence, I exclude banking and insurance (2-digit primary SIC codes 60-67) as well as communication, energy, and transport (40-49). In addition, I exclude companies with a primary SIC code in public administration (91-99). Third, instead of using all firm-fiscal year observations, I use a matching approach where the sample is subdivided in high and low data intensity firms using a median split. To find comparable firms among the high ('treated') and low ('control') data intensity firms, I estimate propensity scores using

a logistic regression of the data firm dummy on all control variables that are measured on the firm-fiscal year level.[11] These propensity scores are then used for one-to-one matching of deals using common support, ties, and a maximum distance of a quarter of the propensity scores' standard deviation. Fourth, instead of applying the tf-idf vectorizer to measure the weighted occurrence of a word in a firm's text, I use a dummy vector indicating whether a word occurs (1) or not (0). See Online Appendix A.1 for more details. Fifth, instead of using the 10-K business descriptions and risk factors of specific firms as the benchmark text for data intensive firms, I use the text from the Economist's "Data Economy Special Issue" (The Economist 2020). Sixth, instead of using a dummy for whether a company acted as an acquirer in a year, I use the natural logarithm of one plus the number of acquisitions within one fiscal year as the dependent variable. Seventh, I further control for a dummy turning to one for the seven case firms. Finally, instead of the sales-based HHI measure provided by Hoberg and Phillips (2010, 2016), I use their product similarity measure to proxy for the extent a firm faces competition. Additional robustness checks for the construction of the data intensity measure are discussed in the following section.

Alternative data intensity measures

The selection of firms used as a benchmark in the measure presented above is based on merger reviews that discussed concerns related to data privacy, access, or aggregation in a possibly combined entity. While this characteristic is central for data firms it can also be restrictive. Using merger reviews to find data intensive firms is directly linked to M&A activity. The benchmark firms include four firms due to their activity in M&A as acquirers, i.e., similar firms might cover typical acquirer characteristics and not necessarily only data firm characteristics. Therefore, in an additional measure, I use all U.S. firms listed in the Nasdaq Yewno Global Artificial Intelli-

[11] These are: Tobin's Q, R&D stock, R&D intensity, No R&D (0/1), tangible, cash, leverage, ROA, sales, PE ratio, age, intangible, and lifecycle.

gence and Big Data Index (NYGBIG®) as benchmark firms. These firms are not in the benchmark group due to their M&A activity and resulting discussions by competition authorities, but due to their data-related patenting behavior and market capitalization (Nasdaq 2021). Here, I estimate the cosine similarities to all listed U.S. firms that are part of the NYGBIG®. The mean similarity to the overall 55 firms then represents an alternative data intensity measure.[12]

In the NYGBIG® several companies are included which do not necessarily collect sensitive consumer data themselves, but are rather specialists on technology, including among others data analytics software, cloud services, and cyber-security. Therefore, in a further measurement, I split these benchmark companies into two groups: those that have consumer interaction, i.e., (at least partly) B2C companies and pure B2B companies.[13] For each of the two groups, I use the same procedure as before to measure data intensity. The correlations between the previously presented data intensity measure and the three measures based on the NYGBIG® companies are between .944 and .996. I use each measure in a separate regression to avoid multicollinearity issues.[14]
I apply an additional validity check to see to what extent the B2B vs. B2C measures capture different firm characteristics within their groups. I randomly select 400 10-K business descriptions and risk factors and manually sort all these firms into data (45) and non-data 10-Ks (355). Among the data firms, I identify 24 B2B and 21 B2C firms using the same definition as for the NYGBIG® firms. Controlling for year-fixed effects, the manually labeled binary variables positively relate to their respective NYGBIG® B2B and B2C similarity measures ($p\text{-}values = 0.00$).[15]

[12] Note that the constituents used here are those part of the index in September 2023.

[13] NYGBIG® B2B: AMD, Ambarella, Aristanetworks, Asana, Cadence Design Systems, Ciena, Cisco, Commvault Systems, Crowdstrike, Dolby Labs, F5, Fastly, Fortinet, HPE, Intel, IBM, Juniper Networks, Micron, MicroStrategy, Motorola Solutions, NetApp, Netscout Systems, Nutanix, Nvidia, Oracle, Palantir, Palo Alto Networks, Pure Storage, Salesforce, ServiceNow, Silicon Labs, Snowflake, Splunk, Synaptics, Synopsys, Tenable, Teradata, UIPath; NYGBIG® B2C: Alphabet/Google, Apple, Microsoft, Adobe, Amazon, Alarm.com, AT&T, Bank of America, Dropbox, eBay, Nortonlifelock/GenDigital, Intuit, Meta/Facebook, Snap, Uber, Verizon, WesternDigital.

[14] Online Appendix A.8 reports all pairwise correlations.

[15] While the B2B dummy can explain a substantial share of the variation of both, the NYGBIG® B2B and B2C variables (added *adj.* R^2: .169 and .145, respectively), the B2C dummy cannot add to the

Acquirer and target likelihood using alternative data intensity measures

Table 2 furthermore presents the regression results using the similarity to NYGBIG® companies instead of the similarity to companies involved in merger review cases (columns (3)-(5) and (8)-(10)). All proxies for data intensity show similar results. The similar results for B2B and B2C companies as well as the high correlation between the different data intensity measures can indicate that all the presented measures capture data-related capabilities. The results thus suggest that data capabilities play a crucial role in driving the positive association between data intensity and M&A likelihood, rather than data resources (alone). This observation aligns with the fact that the B2C companies used as a benchmark also possess well-established data capability skills, as evidenced by their inclusion in NYGBIG® in the first place. Furthermore, all case firms involved in the merger review cases arguably have well-established data capabilities – reinforcing that this characteristic is captured in the data intensity proxy.

To understand which types of deals drive the positive relation between data intensity and acquirer likelihood, I further distinguish between horizontal, vertical, and conglomerate deals.

INSERT TABLE 2 HERE

Conglomerate vs. horizontal vs. vertical deals

The previous results suggest that data intensive firms are heavily investing in M&A. Two of the explanations for increased M&A incentives mentioned in the literature section are related to the deal type. First, classical market power related incentives to increase economies of scale in particular for platform-based businesses would be visible in more acquisitions of industry peers, i.e., suggesting a large number of horizontal deals. Second, forming conglomerates are particularly attractive for data intensive firms if they can lever their data resources and/or capabilities into

explained variance of NYGBIG® B2B, but to NYGBIG® B2C (added *adj.* R^2: .004 and .031, respectively). These results point towards capturing data capabilities in both NYGBIG® measures (to a larger extent in B2B), while data resources only make a difference in the B2C measure.

(seemingly) unrelated industries. To obtain more insights on whether the deals are targeted at increasing market shares (horizontal), integration along the supply chain (vertical), or enlarging the scope of the firm (conglomerate), I distinguish between these deal types for acquirers.

I use the method presented in Fan and Goyal (2006) to identify horizontal, vertical, and conglomerate deals. I obtain the input-output tables from the Bureau of Economic Analysis (BEA) to identify vertical relatedness. Two deal parties are vertically related when the output of one industry needed to produce the total output of the other industry exceeds 1 percent.[16] Deal parties within the same BEA industry that surpass the 1 percent threshold are considered as vertically related as long as their primary SIC code is not the same. A deal is defined as horizontal if deal parties are in the same BEA industry, but vertical relatedness is below 1 percent, or if deal parties share same primary SIC code. Finally, mergers or acquisitions without vertical nor horizontal relatedness of deal parties are considered as conglomerate deals. Using these definitions, 33 percent of deals are classified as horizontal, 22 percent as vertical, and 44 percent as conglomerate.

INSERT TABLE 3 HERE

The results in Table 3 indicate that conglomerate deals are particularly attractive for data intensive firms. Vertical deals are less likely, while there is no considerable effect for horizontal deals. Using the B2B and B2C firms from the NYGBIG® as data firm benchmarks shows similar results. However, excluding B2C firms reduces the marginal effect (from above .2 to .17) for conglomerate deals. In economic terms, a .1 increase in data intensity is associated with a 2.6 percent higher probability of conglomerate deals and is particularly relevant for more consumer-oriented businesses ($p\text{-}value = .00$). Note that the marginal effects of each variable in each re-

[16] The Bureau of Economic Analysis (BEA) provides input-output tables on a fine-grained level (409 industries) for 2007 and 2012. 2007 BEA data is used for deals before 2011; 2012 BEA data is used for deals in 2011 and later. Thomson Reuters Eikon provides 2007 and 2022 NAICS codes for each deal. 2007 NAICS codes are used for all deals since 2007 is closer to 2012, in which the input-output data is measured, than 2022. Only if 2007 NAICS codes are missing, 2022 NAICS codes are used.

gression sum up to zero, i.e., the positive effect for conglomerate deals is accompanied by a negative effect for horizontal ($p\text{-}value = .15$) and vertical deals ($p\text{-}value = .00$).

The results in Table 3 also reveal that horizontal deals are less likely and conglomerate deals more likely when market concentration (HHI) is high. When market concentration is already high, there might be less further concentration potential, marginally lower potential for economies of scale, or higher risk of intervention by competition authorities. The latter issue might be less relevant in conglomerate deals, but also a company's strategy might be redirected towards economies of scope when economies of scale are already largely exploited in a highly concentrated market.

The results are robust to using (1) a 2 percent cutoff and the 409 industry BEA tables[17] as well as (2) a 5 percent cutoff and the less detailed, but annually updated, 71 industry BEA tables[18].

Furthermore, the results hold and are economically similar when going from the deal level to the firm-fiscal year level and adding the non-acquirer observations to the analyses (not depicted). Then, the three acquisition types must be distinguished on a firm-fiscal year level. The dependent variable in the multinomial regression turns to one (two/three), if the largest category among a firm's acquisitions is horizontal (vertical/conglomerate) in the following fiscal year. Using the 409 industry BEA tables and a 1 percent cutoff to define vertical deals, 25 percent (45%/30%) of acquirers mostly invest in vertical (conglomerate/horizontal) deals. This analysis provides further insights, since the previously observed marginal effect of 19.6 percent higher acquirer likelihood for the data intensity variable can be mostly split into 16.0 percent ($p\text{-}value = 0.00$) for conglomerate, 2.5 percent ($p\text{-}value = 0.18$) for horizontal, and 0.9 percent ($p\text{-}value = 0.51$) for vertical

[17] Note that Fan and Goyal (2006) propose a 5% cutoff for a stricter definition of vertical relatedness. This would entail a looser definition of un-relatedness (i.e., neither horizontally nor vertically related) and therefore a vast majority of conglomerate deals. Since I am particularly interested in un-related deals, I chose a 2% cutoff, resulting in 33% horizontal, 17% vertical, and 50% conglomerate deals.

[18] Here, 2007 NAICS codes are used for deals until 2015, 2022 NAICS codes are used for deals in 2016 and later. With this definition, 38% of deals are horizontal, 20% vertical, and 41% conglomerate.

deals. While B2C data intensity shows a very similar result, the marginal effect of B2B data intensity (18.7%) is slightly differently divided between conglomerate (13.1%, $p\text{-}value = 0.00$), horizontal (4.1%, $p\text{-}value = 0.02$), and vertical (1.1%, $p\text{-}value = 0.46$) deals.

These results give important insights to understand acquisition behavior in the data economy. The stable positive effect of the data intensity measures where B2C firms were included as the benchmark firms (data intensity and NYGBIG® B2C data intensity variable) and the slightly lower economic magnitude for the NYGBIG® B2B data intensity variable indicates that the conglomeration trend is visible in the data. The incentive to form a conglomerate is highest for data owners who can expand not only capabilities but also insights from their generated data to (seemingly) un-related industries. Furthermore, the low marginal effect for the likelihood of horizontal deals in particular for those data intensity variables that include B2C companies in their benchmark group signals that market power and network effect arguments cannot explain large parts of the increased incentive for data intensive firms to acquire.

Announcement returns

As explained in the literature review, the expected relation between data intensity and deal performance is less clear. Superior performance would mean that data intensive acquirers are either better in selecting targets or better in generating benefits from their deals. Hence, to understand whether data intensive firms can gain from this increased M&A activity, I furthermore investigate whether data intensity is related to deal performance. One of the most frequently presented performance measures are cumulative abnormal returns (CAR) around deal announcement. They indicate how investors evaluate the synergy potential of a deal. CARs are estimated using a market model event study regression with a one-year calibration period, a separation period of 10 trading days, and a three-day event window around deal announcement [-1,+1]. In the event

study regressions, I control for confounding effects to the 'normal return'-estimation from other 8-K publications during the calibration period. I estimate CARs for all listed acquirers and targets as well as combined market value-weighted CARs for those deals in which both transaction parties are U.S. headquartered and listed.

The sample of firm-fiscal year observations at time t with M&A activity during year $t+1$ is then divided in to high and low data intensity firms using a median split. To apply a matching approach and thus to find comparable firms among the high ('treated') and low ('control') data intensity firms, I estimate propensity scores on the firm-fiscal year level for acquirers and targets separately based on the control variables that are measured on the firm-fiscal year level.[19] These propensity scores are then used for one-to-one matching of deals using common support, ties, and a maximum distance of a quarter of propensity scores' standard deviation.

Testing for differences of matching variables between treatment and control group reveals that some of these variables still show differences with $p\text{-}values < .1$ between groups. Therefore, I run weighted least squares regressions with firm-clustered standard errors where all control variables that showed a significant difference between treatment and control group (with $p\text{-}values < .1$) are again added as controls. Weights are those resulting from propensity score matching to account for the frequency with which the observation is used as a match. In these regressions, I furthermore add Fama French 49 industry and year fixed effects.

Before discussing the relation between data firm variables and announcement returns, deal specific descriptive statistics are presented in Table 4. Average CAR are all positive with 1.0 percent for acquirers (t-test against zero shows a p-value of .00), 23.7 percent for targets ($p\text{-}value = .00$), and 3.0 percent for combined market value-weighted acquirer and target ($p\text{-}value = .00$). Almost

[19] These are: Tobin's Q, R&D stock, R&D intensity, No R&D (0/1), tangible, cash, leverage, ROA, sales, PE ratio, age, intangible, and lifecycle.

all deals are friendly and one quarter of transactions involve a foreign bidder or target. Public to public transactions are to a larger extent horizontal and less often conglomerate deals than if one deal party is not listed (acquirer or target).

INSERT TABLE 4 HERE

Table 5 reports the regression results on the extent to which data intensity is related to deal CAR. The coefficients of both, the data firm dummy using a median split of data intensity and the continuous data intensity variable have rather high p-values. The only variable of interest with a comparably low p-value of .05 is the interaction term of the continuous data intensity variable and horizontal deals, indicating a positive impact of such market power driven deals for highly data intensive acquirers. However, since the median-split data firm dummy coefficient has a p-value of .42 and an alternative definition of data firms using only the top third of data intensive firms as the treatment group shows similar results as well, the claim that horizontal deals show positive announcement returns for data intensive acquirers is – if anything – quite weak. For target investors results look similar, with no significant differences for data intensive targets, independent of the deal type. In the small sample of public to public transactions, the only consistent result between using the continuous data intensity variable and the data firm dummy is the negative coefficient of conglomerate deals for non-data firms, where investors seem to dislike conglomerate deals when data intensity of both acquirer and target is low. Using the data firm dummy also shows a negative coefficient for vertical deals and a positive interaction with conglomerate deals, but these results do not hold when using the continuous data intensity variable.

INSERT TABLE 5 HERE

The results suggest that data intensive firms are not more successful acquirers than less data intensive ones when looking at deal announcement returns. Explanation for this result can be manifold. For example, if targets are relatively small, information on targets is scarce, and data in-

tensive firms buy more often (as the increased deal likelihood and anecdotal evidence on, e.g., Google suggests), the events can be of lower economic magnitude and less surprising for the market. Furthermore, if these acquisitions are targeted towards long-term goals such as innovation, they are potentially more risky and projected benefits may be more distant in the future. This can lead to greater discounting of predicted synergies and thus a reduction of deal-induced changes of market values. Such uncertainties can be reduced throughout and after target integration. Then, post-acquisition periods should reveal these advantages and result in superior long-term performance (e.g., Loughran and Vijh 1997).

Buy-and-hold abnormal returns

While investors of data intensive firms do not seem to benefit from acquisitions at deal announcement, they might benefit over time with a superior long-term stock performance. Therefore, I measure buy-and-hold abnormal returns (BHAR) using the method proposed in Lyon et al. (1999). I sort monthly stock returns of companies listed at NYSE, Nasdaq, and AMEX retrieved from CRSP into 14 portfolios according to their market value. Each of these 14 portfolios is further divided into market-to-book quintiles, generating overall 70 benchmark portfolios. BHARs are estimated from the difference between a firm's multi-year cumulative log-return starting in the month after deal announcement and the corresponding mean portfolio cumulative log-return from the same time period. I use two, three, and four year time windows. BHARs are winsorized (1/99%) to reduce the impact of extreme observations on coefficients. Then I use the same matching procedure and regression design as in the CAR analyses. As a further control variable I add the number of completed deals announced during the BHAR estimation period. I use the natural logarithm of one plus the number of deals in the regressions.

INSERT TABLE 6 HERE

Table 6 reports the descriptive statistics of the most relevant variables in Panel A and regression results for using two-, three-, and four-year BHAR as dependent variables in Panel B. Mean BHARs are close to zero with standard deviations of 0.49 to 0.64. The number of further completed deals announced during the estimation period ranges from zero to up to 127 deals in four years. On average, firms announce more than one deal per year during the estimation periods, with a mean of 2.8 during two years, 3.5 during three years, and 4.2 during four years after deal announcement. The data intensity variable shows a slightly higher mean than in the overall sample (.43 vs. .39), which can be explained by the fact that here only acquirers are investigated and low data intensity firms are more often non-acquirers.

The regression results show that above-median data intensive firms experience a 7-8 percent higher BHAR than below-median data intensive firms after an acquisition announcement. Using the continuous data intensity variable shows consistent results with a 2-3 percent higher BHAR per .1 increase in data intensity ($p\text{-}values < .06$). The number of deals during the BHAR estimation period also shows a positive result ($p\text{-}values = .00$), indicating that firms with more deals experience overall higher long-term stock returns as well.

The positive coefficient of the data firm dummy can be also observed in Figure 2, where above-median data intensive firms show an overall increasing development of long term stock-returns after deal announcement, while below-median data intensive firms show a rather decreasing trend overall. Stock performance on average decreases in the first months after deal announcement for all firms, but for data intensive firms the line starts to recover after approximately eight months and arrives in the positive BHAR zone after 12-18 months. This indicates that uncertainties of the benefits of data intensive firms' acquisitions can be reduced over time and investors adjust their assessments accordingly.

INSERT FIGURE 2 HERE

Robustness tests show similar results, i.e., when using (1) k-nearest neighbor matching with three neighbors instead of one-to-one matching, (2) BHARs without winsorizing (at 1/99%), and (3) adding deal-level controls (i.e., all cash, cross border, public target, acquisition types horizontal/vertical/conglomerate, hostile, and relative deal value). In the next section, I analyze target characteristics to further examine which of the potential motives discussed in the literature section is most likely responsible for the increased M&A activity.

Who buys whom?

Anecdotal evidence and prior research suggests that powerful firms in the data economy buy particularly young and small firms (e.g., Argentesi et al. 2021; Gautier and Lamesch 2021). In my sample, data intensive acquirers more often invest in private or subsidiary and foreign targets. Deal values are less frequently available in Thomson Reuters Eikon, suggesting that deal sizes are smaller and thus deal values are less often reported. Figure 3 depicts these results, where for each data intensity decile, target statistics are documented. While for all levels of data intensity the share of public targets is low, the numbers of non-public target acquisitions are not only absolutely, but also relatively increasing for higher data intensity values. Less deal value availability is most visible in private/subsidiary acquisitions for very high vs. very low data intensity levels, while there is little difference between high and more medium data intensity levels. Independent of target public status, cross border deals are more frequent for data intensive acquirers, with 10-15 percent for low data intensity deciles and around 30 percent for high data intensity acquirers.

The overall picture suggests that data intensive firms are more often acquiring small and foreign targets. This fits into the result that conglomeration is a potential driver for M&A, since small and often young, potentially foreign firms can bring new perspectives, markets, and innova-

tion into the acquiring firm. While the more frequent acquisitions of small and young firms could point to potential killer acquisition arguments, the result that conglomerate deals are more likely than horizontal transactions indicates that these arguments cannot represent the most important drivers of increased acquisition behavior by data intensive firms.

INSERT FIGURE 3 HERE

For acquirers of data intensive targets the picture is less clear. Since targets are public firms, deal values are mostly available. 35-50 percent of acquirers are non-public, but without a pattern for more or less data intensive targets. Cross border transactions show an inverse-U, with very low and very high data intensity targets having less foreign acquirers and medium data intensity targets with more foreign acquirers. Deal values are rather stable over all target data intensity deciles when acquirers are publicly listed, but a comparably high average deal value in the highest data intensity decile when acquirers are private or subsidiary firms. However, deal values in the highest target data intensity deciles show similar absolute values for both acquirer types, indicating that there is no significant difference in deal sizes of highly data intensive targets when public vs. private/subsidiary firms are acquirers.

While the sample reduces substantially when investigating U.S. public to public deals, it provides more target and acquirer characteristics. In particular, I can investigate data intensity levels of both transaction parties and whether targets of data intensive firms are relatively younger.

The correlation of target and acquirer data intensities is .78, while NYGBIG® B2B (B2C) data intensity correlation is .82 (.76), implying that data intensive firms typically acquire other data intensive companies.[20] Top 10 percent B2B data intensive acquirers acquire targets with higher B2B data intensity than B2C data intensity ($p\text{-}value = .00$). For top 10 percent B2C data intensive acquirers, their target's B2B vs. B2C data intensity scores are more similar. Thus, while top

[20] Online Appendix A.10 shows a graphical depiction of data intensities of different deal parties.

B2B data intensive acquirers seem to be comparably more capability building (investing in more data B2B- than in B2C-data intensive intensive firms), highly B2C data intensive acquirers are both capability building and seeking (investing in similarly B2B- and B2C-data intensive firms).

While acquirer age is positively correlated to target data intensity ($correlation = .05$), target age shows a negative correlation with acquirer data intensity ($correlation = -.04$). Hence, older acquirers seem to slightly favor data intensive targets, while data intensive acquirers rather invest in younger targets. The latter observation would point to potential pre-emptive buyout arguments. However, since the magnitude of the correlation is not very high, it again cannot explain the extent of increased acquisition behavior observed in the acquisition likelihood analyses.

DISCUSSION AND CONCLUSION

The present study provides important insights on acquisition behavior by firms with data driven business models. I develop a new measure using textual analysis that captures the extent to which firms have data resources and capabilities (e.g., skills, tools, and/or equipment for data analysis) and show how this data intensity proxy is reflected in industry categorizations, etc. I then use this measure to investigate the relation between data intensity and acquisition behavior.

The findings indicate that data intensive firms are more active in acquisitions with a 2% higher acquirer likelihood per .1 data intensity increase (data intensity ranges from 0 to 1). Data intensive firms are also attractive targets with a 0.6% higher acquisition likelihood per .1 data intensity increase. This increase in M&A activity can be largely attributed to an increase in conglomerate deals (1.6% per .1 data intensity increase). When looking at firms with higher data capability (not necessarily more data resources), horizontal deals are significantly more likely (0.4% per .1 data intensity increase). Targets of data intensive acquirers are on average smaller, more often unlisted, foreign, and younger than targets of less data intensive acquirers.

These results indicate that conglomeration is the most visible driver behind the increased acquisition incentive by data intensive firms. Horizontal deals are visible as well, but to a much smaller extent. Observed target characteristics could support killer acquisition arguments. However, since these would require more horizontal deals as well, they can at most explain a small share of the increased acquisition activity of data intensive firms. Nevertheless, how powerful data intensive acquirers are, might also be represented in conglomerate deals when exploiting competitive advantages in entering new industries. While low entry barriers can be a positive sign for competition, if they are much lower for large tech companies than for start-ups, the incentive to enter for newcomers might still be impacted. How such M&A behavior affects the formation of new companies and the extent of innovative activities overall, should be analyzed in future studies.

While the main drivers of data intensive companies to invest in M&A seem to be clearer, investors initially do not benefit from such deals. However, data intensive acquirers experience superior long-term performance. One explanation could be that targets are often small, young, and innovative firms. Information on such targets are often scarce. Such acquisitions can be seen as comparably more risky and long-term oriented, potentially leading to greater discounting of expected synergies as well as low (and insignificant) effects for data intensive firms. The superior long-term performance of data intensive firms can crystallize when uncertainties are reduced in the years after a deal announcement. Further research is needed to understand these dynamics.

This study is contributing to the scarce empirical literature on data-driven business models. First, the developed measures for identifying the extent to which firms have data resources and capabilities provide opportunities for further analyses of data intensive businesses, not only in an M&A context. Prior studies on data firms or assets use small samples based on surveys (e.g.,

Brynjolfsson et al. 2011), technical skills of employees using LinkedIn data (e.g., Tambe 2014), or job postings in a single industry (e.g., Abis and Veldkamp 2023) to identify data intensity.

Second, investigating M&A activity by data intensive firms for a large number of companies complements analyses in existing studies, which are mostly case-based or provide anecdotal insights on this topic. For instance, Argentesi et al. (2021) investigate acquisitions by Amazon, Facebook, and Google in 2008-2018 and find that their targets are often young and outside the five companies' core businesses. Gautier and Lamesch (2021) investigate deals by Amazon, Apple, Google, Facebook, and Microsoft in 2015-2017 and demonstrate that products of acquired firms are often discontinued, at least under their original name. Latham et al. (2020) analyze acquisitions by Amazon, Apple, Facebook, and Google in 2009-2020 and find that many deals target acquiring capabilities to enter new markets; only a minority of deals satisfy typical characteristics of 'killer acquisitions'. This paper presents large-scale empirical evidence for the importance of conglomeration for data intensive companies. The results suggest that horizontal deals only explain a small share of increased acquisition activity by data intensive firms. This indicates that the argument of killer acquisitions may not be as pronounced as discussed in some literature and policy debates, reinforcing the argumentation and anecdotal evidence presented by Latham et al. (2020).

Third, exploring how acquisition performance relates to data intensity extends the findings in prior literature, which has focused more on other technological resources and capabilities such as R&D and patents (e.g., Ahuja and Katila 2001; King et al. 2008; Sears and Hoetker 2014). This study observes superior long-term performance of data intensive acquirers. The absence of significantly different abnormal returns at deal announcement can point towards more uncertain investments, such as deals with synergies resulting from innovation. Prior literature suggests that such deals exhibit higher uncertainty compared to deals with cost or financial synergies (as dis-

cussed by Rabier 2017). The risk of such operational synergies can be mitigated over time and then become evident in stock returns over the years following the transaction.

This study has important implications. The results suggest that data intensive companies overall have a significantly higher incentive to invest in corporate takeovers. This result is not driven by few big tech companies, but is valid for a vast range of firms active in the data economy. In the long run data intensive firms seem to experience superior performance after conducting acquisitions, which additionally suggests that they can increase their market power over time. Therefore, regulators and policy makers should not only focus on acquisition behavior by the 'big five' tech giants, but also by many other large tech organizations that heavily invest in M&A. Furthermore, the result that targets are rather small, foreign, and young could be supporting evidence for pre-emptive buyouts. However, the result that most of the increase in acquisitions corresponds to an increase in conglomerate deals suggests that the majority of increased acquisition behavior is targeted towards market entry or exploiting economies of scope. While pre-emptive buyouts are already pointed to be typically too small to be in the focus of competition authorities, the result in this study suggests that targets are not even operating in the same market as their acquirers and such deals are thus potentially even less affected by regulation. Not only the difficult question of market definitions, but also the question of exploiting market power when entering new industries should gain more attention.

This study also suggests that data intensive firms benefit from their acquisitions in the long-run. Data intensive firms seem to be superior in selecting targets or are better in exploiting potential benefits later on. Managers of less data intensive firms should consider investing in such resources and capabilities to catch up with their more data intensive peers and improve performance in the long run.

References

Abis, S. and Veldkamp, L. (2023). The changing economics of knowledge production. *Available at SSRN 3570130*.

Ahuja, G. and Katila, R. (2001). Technological acquisitions and the innovation performance of acquiring firms: a longitudinal study. *Strategic Management Journal*, 22(3):197–220.

Amit, R. and Schoemaker, P. J. (1993). Strategic assets and organizational rent. *Strategic Management Journal*, 14(1):33–46.

Argentesi, E., Buccirossi, P., Calvano, E., Duso, T., Marrazzo, A., and Nava, S. (2021). Merger policy in digital markets: an ex post assessment. *Journal of Competition Law & Economics*, 17(1):95–140.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1):99–120.

Becher, D. A., Griffin, T. P., and Nini, G. (2022). Creditor control of corporate acquisitions. *The Review of Financial Studies*, 35(4):1897–1932.

Bena, J. and Li, K. (2014). Corporate innovations and mergers and acquisitions. *The Journal of Finance*, 69(5):1923–1960.

Bergh, D. D., Ketchen Jr, D. J., Orlandi, I., Heugens, P. P., and Boyd, B. K. (2019). Information asymmetry in management research: Past accomplishments and future opportunities. *Journal of Management*, 45(1):122–158.

Betton, S., Eckbo, B. E., and Thorburn, K. S. (2008). Corporate takeovers. *Handbook of Empirical Corporate Finance*, pages 291–429.

Bourreau, M. and De Streel, A. (2019). Digital conglomerates and eu competition policy. *Available at SSRN 3350512*.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in numbers: how does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

Cabral, L. (2023). Big tech acquisitions. *Working Paper, http://luiscabral.net/economics/workingpapers/bigtech%202023%2006.pdf*.

Cai, J., Song, M. H., and Walkling, R. A. (2011). Anticipation, acquisitions, and bidder returns: Industry shocks and the transfer of information across rivals. *The Review of Financial Studies*, 24(7):2242–2285.

Campbell, J., Goldfarb, A., and Tucker, C. (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy*, 24(1):47–73.

Chondrakis, G. (2016). Unique synergies in technology acquisitions. *Research Policy*, 45(9):1873–1889.

Cohen, W. M. and Levinthal, D. A. (1989). Innovation and learning: the two faces of R&D. *The Economic Journal*, 99(397):569–596.

Cohen, W. M. and Levinthal, D. A. (1990). Absorptive capacity: a new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1):128–152.

Corrado, C., Haskel, J., Iommi, M., Jona-Lasinio, C., and Bontadini, F. (2022). Data, digitization, and productivity. Technical report, National Bureau of Economic Research.

Correia-da Silva, J., Jullien, B., Lefouili, Y., and Pinho, J. (2019). Horizontal mergers between multisided platforms: insights from Cournot competition. *Journal of Economics & Management Strategy*, 28(1):109–124.

Cremers, K. M., Nair, V. B., and John, K. (2009). Takeovers and the cross-section of returns. *The Review of Financial Studies*, 22(4):1409–1445.

Cunningham, C., Ederer, F., and Ma, S. (2021). Killer acquisitions. *Journal of Political Economy*, 129(3):649–702.

Dickinson, V. (2011). Cash flow patterns as a proxy for firm life cycle. *The Accounting Review*, 86(6):1969–1994.

Dierickx, I. and Cool, K. (1989). Asset stock accumulation and sustainability of competitive advantage. *Management Science*, 35(12):1504–1511.

Eisenhardt, K. M. and Martin, J. A. (2000). Dynamic capabilities: what are they? *Strategic Management Journal*, 21(10-11):1105–1121.

European Commission (2008). Summary of Commission decision of 11 March 2008 declaring a concentration compatible with the common market and the functioning of the EEA Agreement (Case COMP/M.4731 - Google/DoubleClick), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2008.184.01.0010.01.ENG&toc=OJ:C:2008:184 :TOC.

European Commission (2014). Mergers: Commission approves acquisition of WhatsApp by Facebook, http://europa.eu/rapid/press-release_IP-14-1088_en.htm.

European Commission (2016a). Commission decision of 23/02/2016 declaring a concentration to be compatible with the common market (Case No COMP/M.7813 - SANOFI / GOOGLE / DMI JV) according to Council Regulation (EC) No 139/2004, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1549876772849&uri=CELEX:32016M7813.

European Commission (2016b). Mergers: Commission approves acquisition of LinkedIn by Microsoft, subject to conditions, http://europa.eu/rapid/press-release_IP-16-4284_en.htm.

European Commission (2018a). Commission decision of 19/10/2018 declaring a concentration to be compatible with the common market (case no comp/m.8994 - microsoft / github) according to council regulation (ec) no 139/2004, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018M8994.

European Commission (2018b). Summary of commission decision of 6 september 2018 declaring a concentration compatible with the internal market and the functioning of the eea agreement (case m.8788 â€" apple/shazam), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018M8788(04).

European Commission (2020). Commission decision of 17 december 2020 declaring a concentration compatible with the internal market and the functioning of the eea agreement (case m.9660 â€“ google/fitbit), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021M9660(02).

European Commission (2021). Commission decision of 21/12/2021 declaring a concentration to be compatible with the common market (case no comp/m.10290 - microsoft / nuance) according to council regulation (ec) no 139/2004, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021M10290.

European Commission (2022). Commission decision of 27 january 2022 declaring a concentration compatible with the internal market and the functioning of the eea agreement (case m.10262 â€“ meta (formerly facebook) / kustomer), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022M10262(02).

Fan, J. P. and Goyal, V. K. (2006). On the patterns and wealth effects of vertical mergers. *The Journal of Business*, 79(2):877–902.

Farboodi, M., Mihet, R., Philippon, T., and Veldkamp, L. (2019). Big data and firm dynamics. *AEA Papers and Proceedings*, 109:38–42.

Federal Communications Commission (2020). Federal Communications Commission, https://www.fcc.gov/.

Federal Energy Regulatory Commission (2020). Federal Energy Regulatory Commission, https://www.ferc.gov/.

Federal Reserve Board (2020). Board of Governors of the Federal Reserve System, https://www.federalreserve.gov/.

FTC (2007). Federal Trade Commission closes Google/DoubleClick investigation, https://www.ftc.gov/news-events/press-releases/2007/12/federal-trade-commission-closes-googledoubleclick-investigation.

FTC (2014). FTC notifies Facebook, WhatsApp of privacy obligations in light of proposed acquisition, https://www.ftc.gov/news-events/press-releases/2014/04/ftc-notifies-facebook-whatsapp-privacy-obligations-light-proposed.

Fuller, K., Netter, J., and Stegemoller, M. (2002). What do returns to acquiring firms tell us? evidence from firms that make many acquisitions. *The Journal of Finance*, 57(4):1763–1793.

Gautier, A. and Lamesch, J. (2021). Mergers in the digital economy. *Information Economics and Policy*, 54:100890.

Graebner, M. E., Eisenhardt, K. M., and Roundy, P. T. (2010). Success and failure in technology acquisitions: lessons for buyers and sellers. *Academy of Management Perspectives*, 24(3):73–92.

Hoberg, G. and Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: a text-based analysis. *The Review of Financial Studies*, 23(10):3773–3811.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Kamepalli, S. K., Rajan, R., and Zingales, L. (2022). Kill zone. *National Bureau of Economic Research Working Paper 27146*.

Katz, M. L. (2021). Big tech mergers: Innovation, competition for the market, and the acquisition of emerging competitors. *Information Economics and Policy*, 54:100883.

Katz, M. L. and Shapiro, C. (1985). Network externalities, competition, and compatibility. *American Economic Review*, 75(3):424–440.

Kaul, A. (2012). Technology and corporate scope: Firm and rival innovation as antecedents of corporate transactions. *Strategic Management Journal*, 33(4):347–367.

King, D. R., Slotegraaf, R. J., and Kesner, I. (2008). Performance implications of firm resource interactions in the acquisition of R&D-intensive firms. *Organization Science*, 19(2):327–340.

Kohers, N. and Kohers, T. (2001). Takeovers of technology firms: Expectations vs. reality. *Financial Management*, 30(3):35–54.

Lambrecht, A. and Tucker, C. E. (2017). Can big data protect a firm from competition? *CPI Chronicle, January 2017*.

Latham, O., Tecu, I., and Bagaria, N. (2020). Beyond killer acquisitions: Are there more common potential competition issues in tech deals and how can these be assessed. *CPI Antitrust Chronicle*, 2(2):26–37.

Loughran, T. and McDonald, B. (2023a). Loughran-mcdonald master dictionary w/ sentiment word lists, https://sraf.nd.edu/loughranmcdonald-master-dictionary/.

Loughran, T. and McDonald, B. (2023b). Stage One 10-X parse data, https://sraf.nd.edu/data/stage-one-10-x-parse-data/.

Loughran, T. and McDonald, B. (2023c). Stopwords, https://sraf.nd.edu/textual-analysis/stopwords/.

Loughran, T. and Vijh, A. M. (1997). Do long-term shareholders benefit from corporate acquisitions? *The Journal of Finance*, 52(5):1765–1790.

Lyon, J. D., Barber, B. M., and Tsai, C. L. (1999). Improved Methods for Tests of Long-Run Abnormal Stock Returns. *The Journal of Finance*, 54(1):165–201.

Martynova, M. and Renneboog, L. (2008). A century of corporate takeovers: What have we learned and where do we stand? *Journal of Banking & Finance*, 32(10):2148–2177.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10):60–68.

Motta, M. and Peitz, M. (2021). Big tech mergers. *Information Economics and Policy*, 54:100868.

Mulherin, J. H., Netter, J. M., and Poulsen, A. B. (2017). The evidence on mergers and acquisitions: a historical and modern report. In *The Handbook of the Economics of Corporate Governance*, volume 1, pages 235–290. Elsevier.

Nasdaq (2021). Nasdaq Yewno Global Artificial Intelligence and Big Data Index , https://www.nasdaq.com/articles/nasdaq-yewno-global-artificial-intelligence-and-big-data-index-2021-09-15.

Prado, T. S. and Bauer, J. M. (2022). Big tech platform acquisitions of start-ups and venture capital funding for innovation. *Information Economics and Policy*, 59:100973.

Rabier, M. R. (2017). Acquisition motives and the distribution of acquisition performance. *Strategic Management Journal*, 38(13):2666–2681.

Sears, J. and Hoetker, G. (2014). Technological overlap, technological capabilities, and resource recombination in technological acquisitions. *Strategic Management Journal*, 35(1):48–67.

Surface Transportation Board (2020). Surface Transportation Board, https://www.stb.gov/.

Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science*, 60(6):1452–1469.

The Economist (2020). A deluge of data is giving rise to a new economy, https://www.economist.com/special-report/2020/02/20/a-deluge-of-data-is-giving-rise-to-a-new-economy.

Trautwein, F. (1990). Merger motives and merger prescriptions. *Strategic management journal*, 11(4):283–295.

Uhlenbruck, K., Hitt, M. A., and Semadeni, M. (2006). Market value effects of acquisitions involving internet firms: a resource-based analysis. *Strategic Management Journal*, 27(10):899–913.

Veldkamp, L. (2023). Valuing data as an asset. *Review of Finance*, 27(5):1545–1562.

Wang, L. and Zajac, E. J. (2007). Alliance or acquisition? a dyadic perspective on interfirm resource combinations. *Strategic Management Journal*, 28(13):1291–1317.

Zhu, F. and Liu, Q. (2018). Competing with complementors: An empirical look at amazon. com. *Strategic Management Journal*, 39(10):2618–2642.

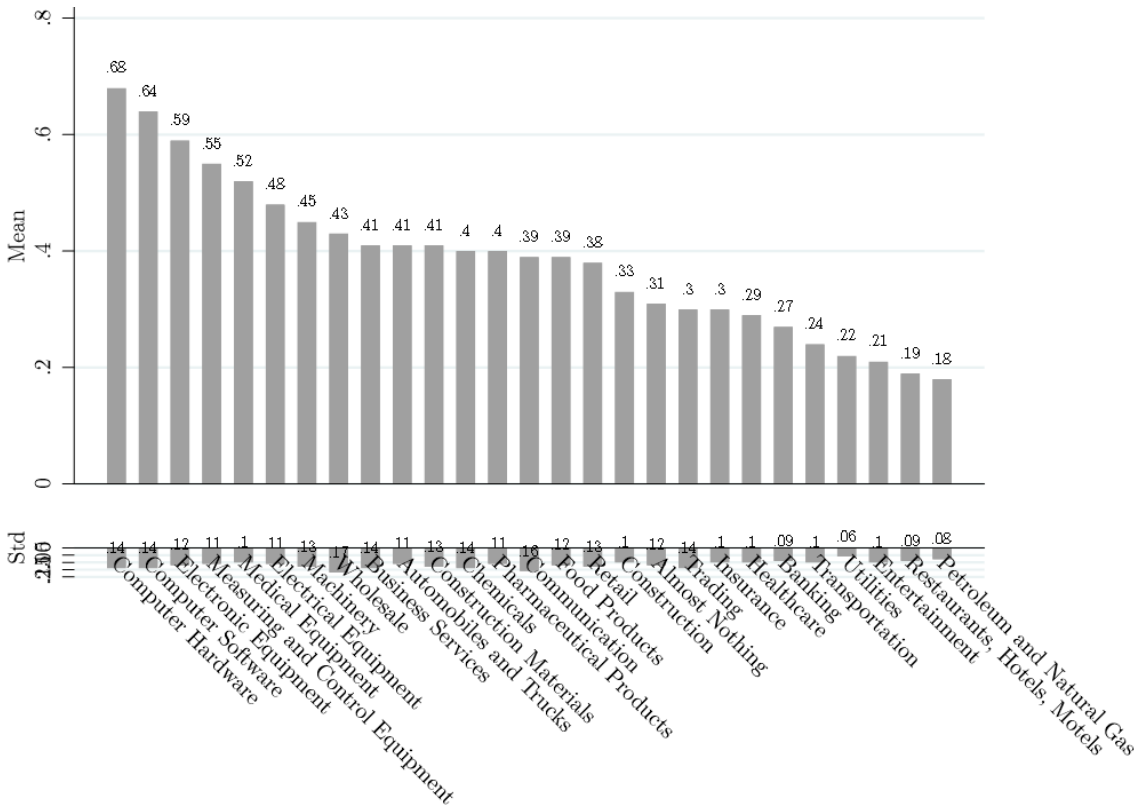

FIGURE 1 Mean and standard deviation of data intensity per Fama French 49 industry

The figure shows the mean and standard deviation of data intensity scores for firms in Fama French 49 industries. Industries are depicted if there are at least 500 observations in the sample.

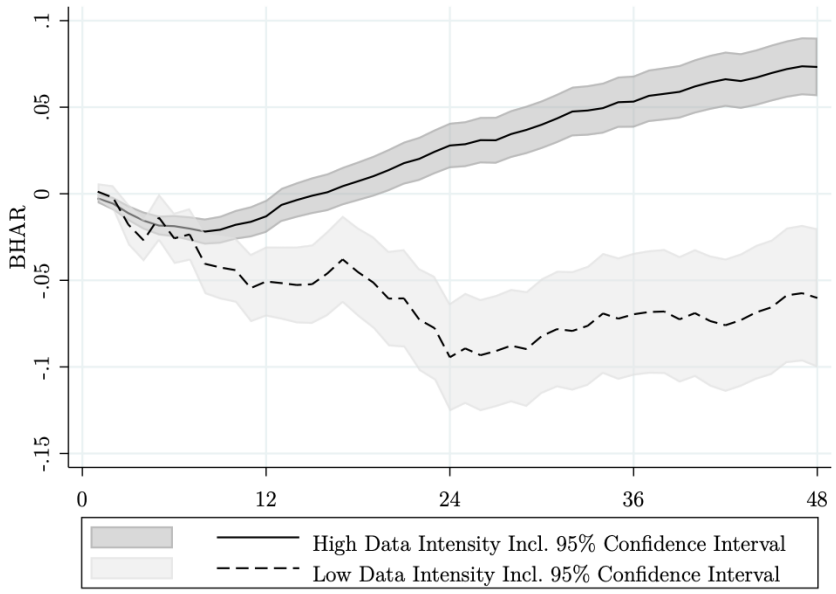

FIGURE 2 Development of BHAR over Four Years After Deal Announcement

The figure depicts the development of buy-and-hold abnormal returns over a period of 48 months for a matched sample of above median and below median data intensive firms. The same weights as in the weighted least squares regression are applied to estimate mean and standard deviations. Grey areas around the mean lines are 95% confidence intervals.

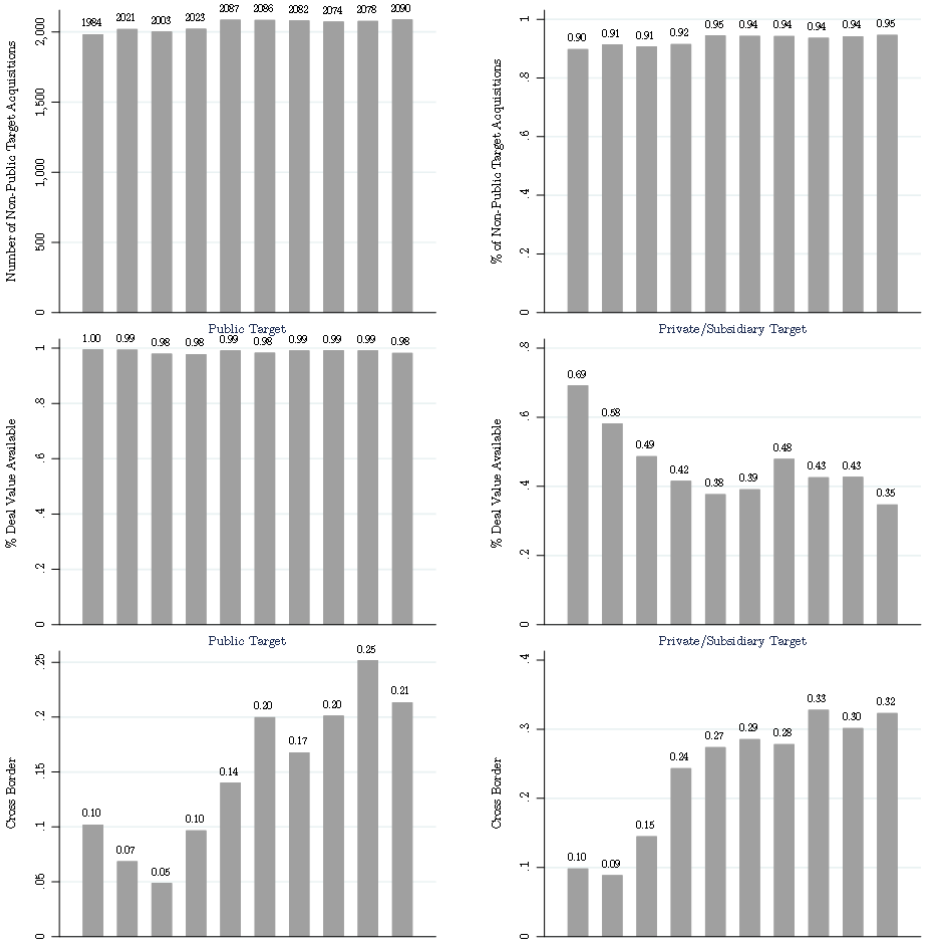

FIGURE 3 Target Characteristics

The figures show observable target characteristics. The figures in the 1st row display the count (top left) and share (top right) of acquisitions of private/subsidiary targets in the sample. The figures in the 2nd row show the share of acquisitions of a public (middle left) and private/subsidiary (middle right) target that provide a deal value in Thomson Reuters Eikon. The figures in the 3rd row report the percentage of cross border deals when acquirers buy public (bottom left) and private/subsidiary (bottom right) targets.

TABLE 1 Descriptive Statistics

This table presents the descriptive statistics and explanations for all control variables. Mean, median, standard deviations (S.D.), minimum/maximum, as well as descriptions and data sources are reported.

Variable	Mean	Median	S.D.	Min	Max	Description	Source
Acquirer (0/1)	0.266	0.000		0.000	1.000	turns to 1 if firm was an acquirer in a majority deal at time t+1	Eikon
Target (0/1)	0.048	0.000		0.000	1.000	turns to 1 if firm was a target in a majority deal at time t+1	Eikon
Data intensity	0.390	0.372	0.177	0.000	1.000	see section on data and methods.	
R&D intensity	0.216	0.000	1.600	-1.229	14.351	3-year average SIC4-industry median adjusted R&D to sales (see King et al. 2008, building on Dierickx and Cool 1989; Cohen and Levinthal 1989, 1990); R&D is set to zero if missing; winsorized (1/99%)	Compustat
R&D stock	1.800	0.000	2.423	-0.436	11.228	3-year depreciated (15% rate) sum of R&D investments (see King et al. 2008); R&D is set to zero if missing; natural logarithm of (1+R&Dstock)	Compustat
No R&D (0/1)	0.483	0.000		0.000	1.000	dummy equal to one if R&D was missing in data	Compustat
Tobin's Q	0.474	0.351	0.621	-5.211	7.891	natural logarithm of the market to book value of firm assets	Compustat
Tangible	0.209	0.114	0.236	0.000	0.894	property, plant, and equipment to total assets; winsorized (1/99%)	Compustat
Cash	0.180	0.097	0.209	0.001	0.967	cash to total assets; winsorized (1/99%)	Compustat
Leverage	0.408	0.360	0.274	-0.204	0.966	1 - market value of equity to market value of firm assets; winsorized (1/99%)	Compustat
ROA	-0.008	0.026	0.191	-1.323	0.315	net income to total assets; winsorized (1/99%)	Compustat
Sales	6.258	6.308	2.100	0.002	13.230	natural logarithm of firm sales as firm size indicator	Compustat
HHI	0.276	0.163	0.270	0.015	1.000	sales based Herfindahl Index within firm's TNIC-3 industry network (Hoberg and Phillips 2010, 2016)	
PE ratio	14.174	14.389	59.299	-284.013	334.123	market value of equity to net income; winsorized (1/99%)	Compustat
Age	23.734	19.000	15.815	2.000	71.000	Compustat firm age, calculated from the observation year minus first Compustat listing year	Compustat
Intangible	0.061	0.019	0.092	0.000	0.473	other intangible assets (i.e., excl. goodwill, etc.) to total assets; winsorized (1/99%)	Compustat
Lifecycle (1/5)	2.769	3.000	0.966	1.000	5.000	Dickinson (2011)'s life cycle measure attributes the stages *Introduction* (1), *Growth* (2), *Mature* (3), *Shake-Out* (4), and *Decline* (5) to firm-fiscal year observations according to the signs of cash flows from operating, financing, and investing activities	Compustat
N	37,760						

TABLE 2 Logistic Regressions for Acquirer and Target Likelihood

This table presents the logistic regressions using an acquirer (target) dummy turning to one for companies that become an acquirer (target) in the following fiscal year as the dependent variable. FF49 industry and year fixed effects are interacted to control for industry specific merger waves. Note that these fixed effects are responsible for the lower number of observations compared to the descriptive statistics in Table 1. Several industry-years are omitted since there are industry-years in which no deal occurred (thus 100% collinear to the constant). Standard errors are clustered on the firm level. Below the marginal effects, p-values are reported.

	Acquirer					Target				
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Data intensity		0.196					0.057			
		[0.000]					[0.000]			
NYGBIG® intensity			0.187					0.057		
			[0.000]					[0.000]		
NYGBIG® B2C				0.183					0.055	
				[0.000]					[0.000]	
NYGBIG® B2B					0.187					0.057
					[0.000]					[0.000]
R&D Intensity	0.002	0.004	0.004	0.004	0.004	-0.004	-0.003	-0.003	-0.003	-0.003
	[0.515]	[0.279]	[0.285]	[0.292]	[0.289]	[0.001]	[0.003]	[0.003]	[0.002]	[0.003]
R&D Stock	0.014	0.011	0.011	0.011	0.011	0.003	0.002	0.002	0.002	0.002
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.002]	[0.036]	[0.042]	[0.026]	[0.046]
No R&D	0.035	0.038	0.037	0.036	0.037	0.005	0.006	0.006	0.005	0.006
	[0.002]	[0.001]	[0.001]	[0.002]	[0.001]	[0.294]	[0.214]	[0.230]	[0.243]	[0.228]
Tobin's Q	0.011	0.012	0.012	0.012	0.012	-0.016	-0.016	-0.016	-0.016	-0.016
	[0.161]	[0.109]	[0.110]	[0.113]	[0.110]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Tangible	-0.196	-0.166	-0.168	-0.167	-0.169	-0.019	-0.010	-0.010	-0.010	-0.011
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.039]	[0.278]	[0.262]	[0.264]	[0.246]
Cash	-0.148	-0.150	-0.148	-0.148	-0.149	-0.018	-0.017	-0.016	-0.016	-0.016
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.050]	[0.074]	[0.085]	[0.084]	[0.083]
Leverage	-0.180	-0.170	-0.172	-0.174	-0.171	0.039	0.041	0.041	0.041	0.041
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
ROA	0.136	0.146	0.145	0.146	0.144	-0.005	-0.004	-0.004	-0.004	-0.004
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.507]	[0.643]	[0.604]	[0.638]	[0.585]
Sales	0.054	0.053	0.053	0.052	0.053	-0.008	-0.008	-0.008	-0.008	-0.008
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
HHI	0.010	0.004	0.005	0.004	0.006	-0.025	-0.026	-0.026	-0.026	-0.026
	[0.461]	[0.726]	[0.684]	[0.735]	[0.654]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
P/E Ratio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	[0.118]	[0.127]	[0.126]	[0.129]	[0.125]	[0.577]	[0.581]	[0.585]	[0.586]	[0.584]
Age	-0.001	-0.000	-0.000	-0.001	-0.000	-0.001	-0.001	-0.001	-0.001	-0.001
	[0.012]	[0.042]	[0.031]	[0.027]	[0.032]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Intangible	0.226	0.214	0.220	0.220	0.221	-0.010	-0.009	-0.007	-0.007	-0.007
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.555]	[0.576]	[0.683]	[0.664]	[0.686]
Lifecycle stage: growth	0.052	0.049	0.049	0.049	0.049	0.012	0.011	0.011	0.011	0.011
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.012]	[0.015]	[0.016]	[0.016]	[0.016]
Lifecycle stage: mature	0.010	0.007	0.007	0.007	0.007	0.018	0.018	0.018	0.018	0.018
	[0.401]	[0.567]	[0.562]	[0.561]	[0.556]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Lifecycle stage: shake-out	-0.033	-0.034	-0.033	-0.034	-0.033	0.023	0.023	0.023	0.023	0.023
	[0.008]	[0.007]	[0.008]	[0.007]	[0.008]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Lifecycle stage: decline	-0.022	-0.019	-0.019	-0.020	-0.019	0.005	0.006	0.006	0.006	0.006
	[0.145]	[0.203]	[0.209]	[0.196]	[0.212]	[0.343]	[0.277]	[0.274]	[0.283]	[0.273]
Constant	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
FF49 X year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	37,556	37,556	37,556	37,556	37,556	32,794	32,794	32,794	32,794	32,794
Pseudo R^2	0.143	0.145	0.145	0.145	0.145	0.0614	0.0631	0.0634	0.0633	0.0633
Pearson chi^2	38948	39049	39048	39028	39053	33327	33295	33284	33353	33253

TABLE 3 Multinomial Logistic Regressions for the Likelihood of Horizontal vs. Vertical vs. Conglomerate Deals

This table presents multinomial logistic regressions using a categorical variable distinguishing between horizontal (H), vertical (V), and conglomerate (C) acquisitions. Standard errors are clustered on the firm level. FF49 industry and year fixed effects are included to control for industry specific and macroeconomic confounders. Below marginal effects, p-values are reported.

Variable	**(1)**			**(2)**			**(3)**		
	H	*V*	*C*	*H*	*V*	*C*	*H*	*V*	*C*
Data intensity	-0.084	-0.180	0.264						
	[0.149]	[0.000]	[0.000]						
NYGBIG® B2C				-0.084	-0.133	0.217			
				[0.115]	[0.001]	[0.000]			
NYGBIG® B2B							0.020	-0.192	0.172
							[0.714]	[0.000]	[0.001]
R&D intensity	-0.009	0.001	0.008	-0.009	0.002	0.007	-0.008	0.001	0.007
	[0.113]	[0.792]	[0.238]	[0.114]	[0.727]	[0.270]	[0.161]	[0.806]	[0.303]
R&D stock	0.017	-0.014	-0.003	0.017	-0.015	-0.002	0.015	-0.014	-0.001
	[0.004]	[0.000]	[0.644]	[0.005]	[0.000]	[0.786]	[0.008]	[0.000]	[0.838]
No R&D	0.060	-0.054	-0.006	0.060	-0.053	-0.007	0.063	-0.055	-0.008
	[0.020]	[0.003]	[0.815]	[0.020]	[0.003]	[0.782]	[0.016]	[0.003]	[0.739]
Tobin's Q	0.011	-0.041	0.031	0.011	-0.041	0.031	0.010	-0.041	0.031
	[0.461]	[0.000]	[0.031]	[0.461]	[0.000]	[0.031]	[0.489]	[0.000]	[0.028]
Tangible	0.203	-0.114	-0.090	0.203	-0.107	-0.097	0.216	-0.111	-0.105
	[0.000]	[0.012]	[0.077]	[0.000]	[0.018]	[0.056]	[0.000]	[0.014]	[0.037]
Cash	0.085	-0.035	-0.050	0.085	-0.037	-0.048	0.081	-0.037	-0.044
	[0.091]	[0.303]	[0.339]	[0.091]	[0.267]	[0.364]	[0.107]	[0.271]	[0.402]
Leverage	0.053	-0.031	-0.022	0.054	-0.025	-0.028	0.057	-0.030	-0.027
	[0.187]	[0.329]	[0.612]	[0.182]	[0.421]	[0.525]	[0.159]	[0.336]	[0.548]
ROA	0.050	-0.005	-0.045	0.050	-0.004	-0.047	0.055	-0.004	-0.051
	[0.250]	[0.911]	[0.347]	[0.249]	[0.932]	[0.334]	[0.211]	[0.921]	[0.294]
Sales	-0.025	0.003	0.022	-0.024	0.003	0.021	-0.025	0.003	0.022
	[0.000]	[0.436]	[0.000]	[0.000]	[0.380]	[0.000]	[0.000]	[0.458]	[0.000]
HHI	-0.109	-0.029	0.138	-0.108	-0.030	0.138	-0.114	-0.029	0.143
	[0.000]	[0.150]	[0.000]	[0.000]	[0.142]	[0.000]	[0.000]	[0.148]	[0.000]
PE ratio	-0.000	0.000	0.000	-0.000	0.000	0.000	-0.000	0.000	0.000
	[0.086]	[0.412]	[0.438]	[0.083]	[0.439]	[0.408]	[0.075]	[0.434]	[0.389]
Age	-0.001	0.001	0.000	-0.001	0.001	0.000	-0.001	0.001	0.000
	[0.095]	[0.052]	[0.746]	[0.103]	[0.037]	[0.851]	[0.117]	[0.040]	[0.873]
Intangible	0.167	-0.220	0.053	0.167	-0.225	0.058	0.160	-0.224	0.064
	[0.039]	[0.001]	[0.530]	[0.038]	[0.001]	[0.501]	[0.049]	[0.001]	[0.450]
Lifecycle stage: growth	0.047	0.026	-0.072	0.046	0.025	-0.071	0.044	0.026	-0.070
	[0.034]	[0.146]	[0.004]	[0.036]	[0.165]	[0.005]	[0.046]	[0.145]	[0.006]
Lifecycle stage: mature	0.028	0.029	-0.058	0.028	0.028	-0.056	0.025	0.030	-0.055
	[0.204]	[0.082]	[0.019]	[0.212]	[0.096]	[0.022]	[0.269]	[0.079]	[0.027]
Lifecycle stage: shake-out	0.012	0.024	-0.036	0.012	0.023	-0.035	0.009	0.024	-0.033
	[0.624]	[0.194]	[0.178]	[0.635]	[0.215]	[0.195]	[0.723]	[0.194]	[0.222]
Lifecycle stage: decline	-0.016	0.035	-0.019	-0.016	0.035	-0.019	-0.015	0.033	-0.018
	[0.560]	[0.151]	[0.540]	[0.549]	[0.150]	[0.549]	[0.585]	[0.169]	[0.564]
Constant	yes	yes	yes	yes	yes	yes	yes	yes	yes
FF49 & year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
N	19,493			19,493			19,493		
Pseudo R^2	0.120			0.119			0.119		

TABLE 4 Descriptive Statistics Deals

This table presents the descriptive statistics for all deal related variables. Mean, standard deviations, minimum and maximum are reported. Summary statistics are weighted according to the propensity score matching. CAR are cumulative abnormal returns around deal announcement [-1/+1]. Relative deal size refers to the natural logarithm of deal value to acquirer market value four weeks prior deal announcement. Hostile (0/1) is a dummy turning to one if Eikon lists the deal as hostile. Toehold (0/1) is a dummy turning to one if the acquirer held at least 5% of the target's shares before the deal. All cash (0/1) is a dummy turning to one if the deal is paid 100% cash. Cross border is a dummy turning to one if the target or acquirer nation is not the United States. Public target/acquirer (0/1) are dummies turning to one if the target/acquirer are publicly listed. Vertical, Horizontal, and Conglomerate are defined using BEA 409 industry tables and a 1% cutoff for vertical relatedness.

	Acquirer				Target				Acquirer + Target			
Variable	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max
CAR	0.01	0.07	-0.49	0.80	0.24	0.21	-0.57	1.59	0.03	0.07	-0.32	0.39
Rel. deal size	-3.12	1.74	-11.2	5.02					-2.19	1.71	-8.29	1.07
Hostile (0/1)	0.00		0.00	1.00	0.00		0.00	1.00	0.00		0.00	0.00
Toehold (0/1)	0.01		0.00	1.00	0.04		0.00	1.00	0.01		0.00	1.00
All cash (0/1)	0.46		0.00	1.00	0.71		0.00	1.00	0.63		0.00	1.00
Cross border (0/1)	0.25		0.00	1.00	0.27		0.00	1.00				
Public target (0/1)	0.16		0.00	1.00								
Public acquirer (0/1)					0.62		0.00	1.00				
Vertical (0/1)	0.20		0.00	1.00	0.18		0.00	1.00	0.17		0.00	1.00
Horizontal (0/1)	0.35		0.00	1.00	0.32		0.00	1.00	0.52		0.00	1.00
Conglomerate (0/1)	0.46		0.00	1.00	0.50		0.00	1.00	0.31		0.00	1.00
N	3,732				1,054				285			

TABLE 5 Weighted Least Squares Regressions for Analyzing CAR

This table presents the weighted least squares regressions using cumulative abnormal returns at deal announcement as the dependent variable and weights from propensity score matching. Standard errors are clustered on the firm-level. Columns (1)-(4) show regressions using acquirer CAR, (5)-(8) using target CAR, (9)-(12) using market value-weighted combined target and acquirer CAR as the dependent variable. Columns (9)-(12) report coefficients of each, target and acquirer firm characteristics next to each other, while coefficients of deal variables are centered between them. Vertical deals are the base category where Horizontal and Conglomerate deal coefficients are shown. Below the coefficients, p-values are reported.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**		**(10)**		**(11)**		**(12)**	
Dep. Variable	Acquirer CAR				Target CAR				Combined Acquirer (A) & Target (T) CAR							
									A	**T**	**A**	**T**	**A**	**T**	**A**	**T**
Data firm dummy	-0.002	-0.004			0.011	0.019			-0.012	-0.010	-0.027	-0.054				
	[0.604]	[0.574]			[0.513]	[0.571]			[0.436]	[0.560]	[0.271]	[0.048]				
Data intensity			0.014	-0.015			0.036	-0.011					-0.053	-0.008	-0.118	-0.065
			[0.364]	[0.499]			[0.610]	[0.934]					[0.273]	[0.914]	[0.142]	[0.453]
Horizontal	0.004	0.001	0.005	-0.017	-0.031	-0.043	-0.031	-0.080	0.017		-0.030		0.018		-0.060	
	[0.468]	[0.947]	[0.439]	[0.248]	[0.170]	[0.241]	[0.171]	[0.336]	[0.156]		[0.269]		[0.145]		[0.211]	
Conglomerate	0.006	0.008	0.006	-0.001	-0.014	0.004	-0.014	-0.024	-0.002		-0.091		-0.004		-0.105	
	[0.139]	[0.221]	[0.147]	[0.926]	[0.518]	[0.901]	[0.518]	[0.751]	[0.898]		[0.009]		[0.741]		[0.080]	
Data firm dummy		0.009				0.024					0.021	0.043				
X horizontal		[0.418]				[0.573]					[0.486]	[0.107]				
Data firm dummy		-0.003				-0.029					0.019	0.093				
X conglomerate		[0.682]				[0.472]					[0.586]	[0.010]				
Data intensity				0.048				0.110							0.029	0.125
X horizontal				[0.049]				[0.481]							[0.767]	[0.225]
Data intensity				0.018				0.027							0.124	0.067
X conglomerate				[0.411]				[0.846]							[0.201]	[0.509]
Rel. deal size	0.005	0.005	0.005	0.005					0.022		0.022		0.023		0.024	
	[0.000]	[0.000]	[0.000]	[0.000]					[0.013]		[0.014]		[0.013]		[0.011]	
Hostile	0.017	0.015	0.017	0.016	0.202	0.186	0.200	0.197								
	[0.043]	[0.073]	[0.045]	[0.056]	[0.000]	[0.000]	[0.000]	[0.000]								
Toehold	0.005	0.006	0.006	0.006	-0.083	-0.080	-0.083	-0.081	0.072		0.082		0.069		0.086	
	[0.485]	[0.432]	[0.467]	[0.442]	[0.344]	[0.343]	[0.343]	[0.345]	[0.095]		[0.063]		[0.089]		[0.057]	
All cash	0.012	0.011	0.012	0.011	0.079	0.077	0.079	0.078	0.022		0.021		0.021		0.019	
	[0.002]	[0.002]	[0.002]	[0.002]	[0.000]	[0.000]	[0.000]	[0.000]	[0.019]		[0.026]		[0.021]		[0.040]	
Cross border	-0.006	-0.006	-0.006	-0.006	0.001	-0.000	0.002	0.000								
	[0.182]	[0.183]	[0.170]	[0.177]	[0.939]	[0.999]	[0.929]	[0.984]								
Public target	-0.024	-0.024	-0.024	-0.024												
	[0.000]	[0.000]	[0.000]	[0.000]												
Public acquirer					0.045	0.046	0.045	0.046								
					[0.015]	[0.011]	[0.015]	[0.014]								
Matching controls	yes	yes	yes	yes	yes	yes	yes	yes	yes		yes		yes		yes	
Constant	yes	yes	yes	yes	yes	yes	yes	yes	yes		yes		yes		yes	
FF49 & year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes		yes		yes		yes	
Observations	3,743	3,743	3,743	3,743	1,054	1,054	1,054	1,054	285		285		285		285	
Adjusted R^2	0.085	0.086	0.086	0.087	0.169	0.170	0.169	0.168	0.482		0.500		0.482		0.485	

TABLE 6 Weighted Least Squares Regressions for Analyzing BHAR

Panel A in this table presents the descriptive statistics of the variables depicted in Panel B. Panel B shows the results of the weighted least squares regressions using buy-and-hold abnormal returns starting in the calendar month after deal announcement as the dependent variable. Time periods used are two (columns(1)-(2)), three (columns(3)-(4)), and four years (columns(5)-(6)). Note that the descriptive statistics in Panel A show the absolute number of deals during the BHAR estimation period. In the regressions in Panel B, $ln(1 + \#Deals\ during\ BHAR\ period)$ are used. Standard errors are clustered on the firm-level. Below the coefficients, p-values are reported.

Panel A: Descriptive Statistics

Variable	Mean	S.D.	Min	Max
3-year BHAR	-0.004	0.566	-1.993	1.351
2-year BHAR	-0.027	0.488	-1.651	1.171
4-year BHAR	0.012	0.644	-2.205	1.509
#Deals during BHAR period				
3-years	3.509	4.339	0	98
2-years	2.773	3.241	0	67
4-years	4.170	5.361	0	127
Data firm dummy	0.500		0.000	1.000
Data intensity	0.425	0.171	0.014	0.936

Panel B: Regression Results

Dependent Variable	2-YEAR BHAR		3-YEAR BHAR		4-YEAR BHAR	
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Data firm dummy	0.074		0.081		0.079	
	[0.016]		[0.005]		[0.046]	
Data intensity		0.197		0.218		0.295
		[0.060]		[0.023]		[0.043]
#Deals during BHAR period	0.075	0.078	0.064	0.067	0.109	0.109
	[0.001]	[0.000]	[0.001]	[0.001]	[0.000]	[0.000]
Matching controls	yes	yes	yes	yes	yes	yes
Constant	yes	yes	yes	yes	yes	yes
FF49 & year FE	yes	yes	yes	yes	yes	yes
Observations	6,184	6,184	6,184	6,184	6,184	6,184
Adjusted R^2	0.184	0.183	0.168	0.167	0.192	0.193

A ONLINE APPENDIX

A.1 Measuring Cosine Similarities

Table 7 shows the text cleaning and cosine similarity estimation steps.

INSERT TABLE 7 HERE

A.2 Relevance of Data Found in Merger Review by Competition Authorities

Table 8 shows text excerpts from the merger review case descriptions.

INSERT TABLE 8 HERE

A.3 Density Histogram Data Intensity

Figure 4 shows the distribution of the normalized data intensity measure.

INSERT FIGURE 4 HERE

A.4 Top 10 Data Intensity Firms in Final Year of the Sample (2021)

In Table 9 I present the data intensity measure of the top 10 scoring data intensity firms in the fiscal year 2021.

INSERT TABLE 9 HERE

A.5 All U.S. firms listed in the NYGBIG® index

In Table 10 I present the data intensity measure for all firms in my sample that are listed in the NYGBIG® index.

INSERT TABLE 10 HERE

A.6 Word Clouds Top and Bottom 100 Data Intensive Firms

I present word clouds for the top 100 and bottom 100 companies sorted according to their data intensity levels in 2015 in Figure 5. I use 2015, since all case firms are included in the sample in that year. The industry distribution as reported in Figure 1 is represented in the word clouds as well. While both word clouds contain words referring to typical 10-K content such as financial results (e.g., operating result, financial condition), the upper word cloud (based on top 100 firms' 10-K Item 1 and 1A text) shows more technology related words including platform, data, network, user, software, etc. The lower word cloud (bottom 100 firms) show more industry specific words as expected from Figure 1, such as natural gas, hotel, oil, vessel, etc.

INSERT FIGURE 5 HERE

A.7 Correlations between dependent, explanatory, and control variables

Table 11 reports all pairwise correlations between the variables presented in Table 1.

INSERT TABLE 11 HERE

A.8 Correlations between Data Intensity Measures

Table 12 reports the pairwise correlations between the previously presented data intensity measure and the three measures based on the NYGBIG® companies.

INSERT TABLE 12 HERE

A.9 Deal values when acquirers are public vs. private/subsidiary firms
Figure 6 shows deal values distinguished between public vs. private/subsidiary acquirers.

INSERT FIGURE 6 HERE

A.10 Data Intensity Levels of Transaction Parties in Public to Public Deals
Figure 7 shows data intensity levels for acquirers and targets in public to public U.S. transactions.

INSERT FIGURE 7 HERE

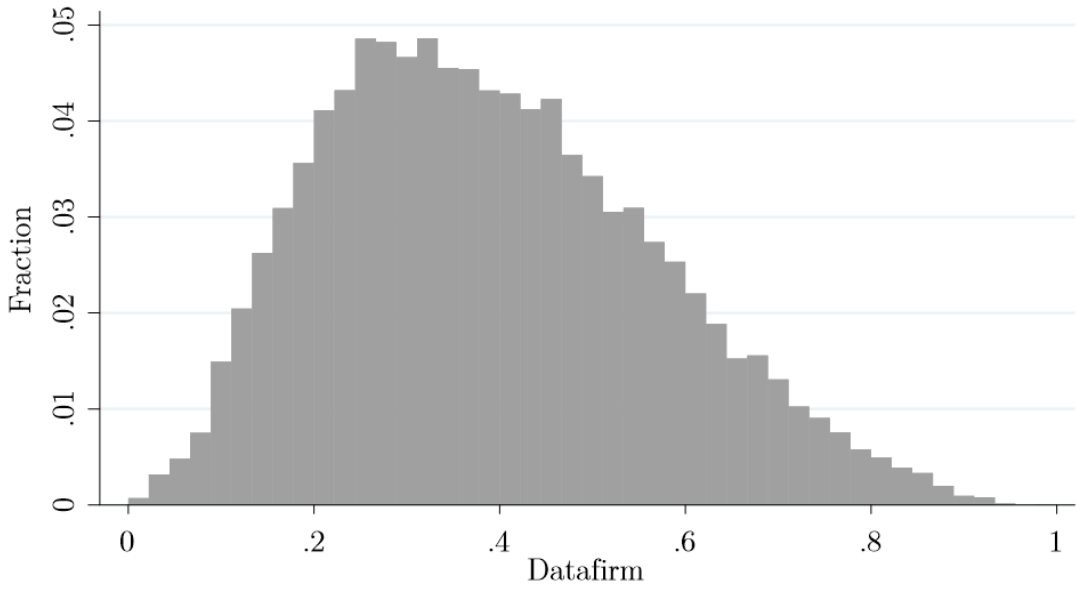

FIGURE 4 Density Histogram Data Intensity Measure

The figure shows a density histogram of the normalized data intensity measure.

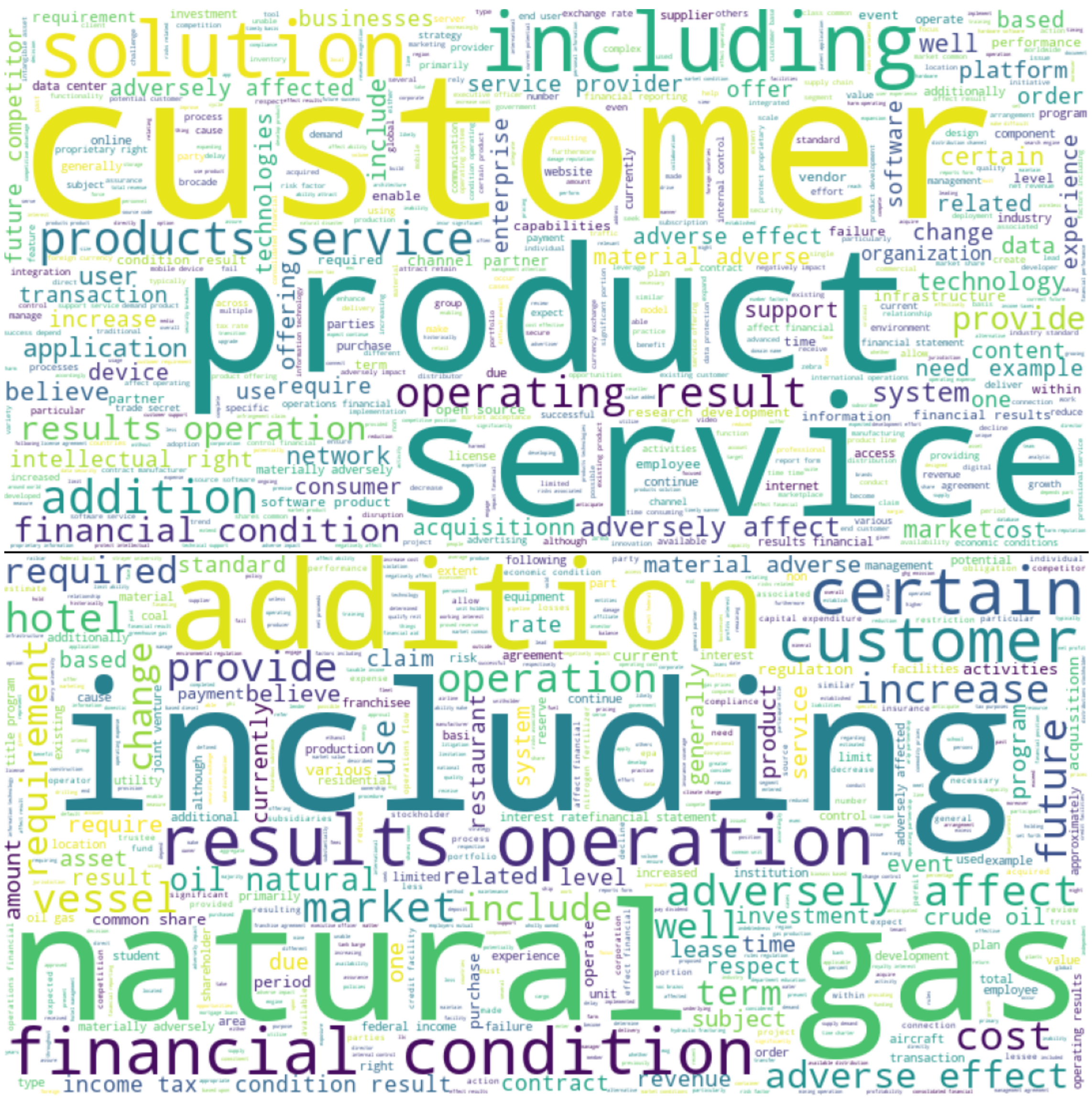

FIGURE 5 Word Clouds for High and Low Data Intensity Scoring Firms

The figure shows word clouds based on the top (upper graph) and bottom (lower graph, divided by the horizontal line) 100 data intensity scoring firms in the fiscal year 2015. 2015 is chosen since all seven case firms are included in the sample in that year.

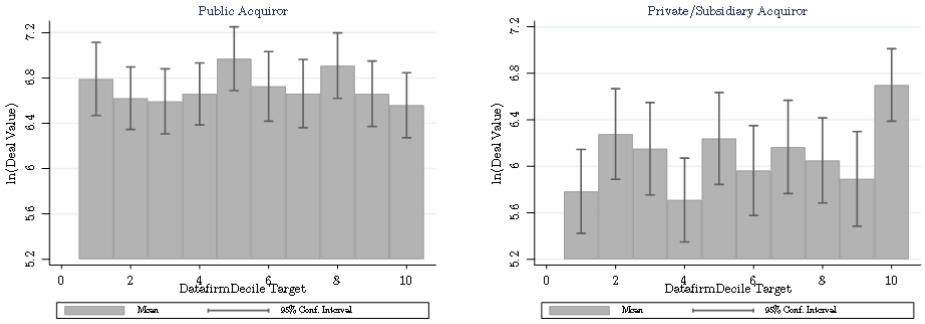

FIGURE 6 Acquirer Characteristics

The figures show deal values of deals including public. vs. private/subsidiary acquirers. The left figure presents the average logarithmized deal values of transactions with public acquirers per target data intensity decile. The right figure depicts the average logarithmized deal values of transactions with private/subsidiary acquirers per target data intensity decile.

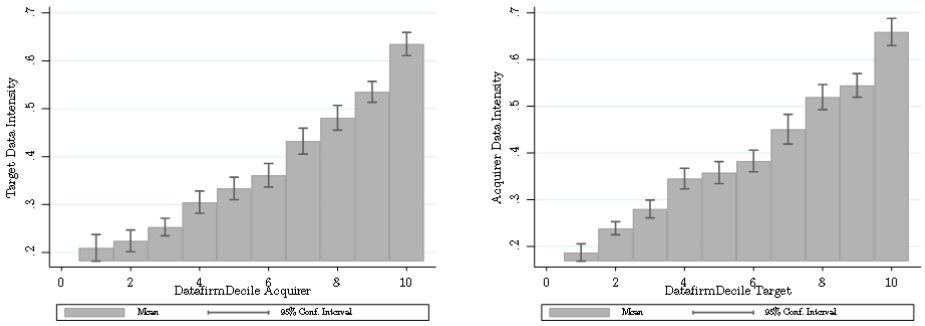

FIGURE 7 Data Intensity Levels of Transaction Parties

The figures show data intensity levels for acquirers and targets in public to public U.S. transactions. The left figure presents the average data intensity levels of targets per acquirer data intensity decile. The right figure depicts the average data intensity levels of acquirers per target data intensity decile. Using the NYGBIG® B2B and B2C distinction shows a very similar picture.

TABLE 7 Text Cleaning and Cosine Similarity Estimation Steps

This table presents the steps towards estimating cosine similarities of item 1 and item 1A in 10-K annual reports. The method is applied for each fiscal year separately.

	Description
1	*Text cleaning*
1a	Including only vocabulary in Loughran and McDonald (2023a) master dictionary.
1b	Excluding stopwords from Loughran and McDonald (2023c) (generic, dates and numbers, geographic, and names).
1c	Using one-grams (no combination of words).
1d	Excluding tokens that appear in less than 1% and more than 99% of all 10-Ks.
1	Resulting number of tokens: between 6,200 and 7,500 tokens for 10-K Business Descriptions (Item 1) and between 4,000 and 5,900 tokens for 10-K Risk Factors (Item IA).
2	*Cosine similarity estimation*
2a	Applying a tf-idf vectorizer to weigh tokens taking into account how often the word appears in all documents (term frequency X inverse document frequency). Resulting vector $P_{i,t}$ for each document: weighted occurrence of tokens within a text (i.e., text '$Item$' = item 1 and item $1A$ respectively for each firm-fiscal year observation). *For robustness checks*: Generate a dummy vector $P_{i,t}$ indicating whether a token occurs in a text (1) or not (0) instead of tf-idf.
2b	Estimating pairwise cosine similarity by applying the following formula to the frequency vectors $P_{i,t}$ and $P_{j,t}$ of two documents i and j in fiscal year t: $SimilarityItem_{ij,t} = P_{i,t}^{'} P_{j,t} (P_{i,t}^{'} P_{i,t})^{-0.5} (P_{j,t}^{'} P_{j,t})^{-0.5}$, which can essentially be interpreted as the pairwise correlation between two texts.

TABLE 8 Text Excerpts from Merger Review Case Descriptions

The table reports excerpts from press releases or case decision texts from the EC or the FTC on the merger cases in which concerns regarding data privacy, access, or aggregation were discussed.

Case	Description
Google/DoubleClick	"6.3.3. Foreclosure based on combination of Google and DoubleClick's assets 25. Finally, the mere combination of DoubleClick's assets with Google's assets, and in particular the databases that both companies have and could develop on customer online behaviour could allow the merged entity to achieve a position that could not be replicated by its competitors. As a result of this combination, Google's competitors would be progressively marginalised which would ultimately allow Google to raise the prices for its intermediation services." (European Commission 2008) "Further, the evidence demonstrated that any aggregation of consumer and competitive data resulting from the acquisition is unlikely to harm competition in the ad intermediation market." (FTC 2007)
Facebook/WhatsApp	p. 29: "5.3. Online advertising services; 5.3.1. Introduction (164) For the purposes of this decision, the Commission has analysed potential data concentration only to the extent that it is likely to strengthen Facebook's position in the online advertising market or in any sub-segments thereof. Any privacy-related concerns flowing from the increased concentration of data within the control of Facebook as a result of the Transaction do not fall within the scope of the EU competition law rules but within the scope of the EU data protection rules." (European Commission 2014) "The letter notes that before making any material changes to how they use data already collected from WhatsApp subscribers, the companies must get affirmative consent. In addition, the letter notes that the companies must not misrepresent the extent to which they maintain the privacy or security of user data. The letter also recommends that consumers be given the opportunity to opt out of any future changes to how newly-collected data is used." (FTC 2014)
Sanofi/Google/DMI	[...] "For the purposes of this decision, the Commission notes that any privacy-related concerns flowing from the use of data within the control of the Parties do not fall within the scope of the EU competition law rules but within the scope of the EU data protection rules." (European Commission 2016a)
Microsoft/LinkedIn	pp. 34f.: "(179) Assuming such data combination is allowed under the applicable data protection legislation, there are two main ways in which a merger may raise horizontal issues as a result of the combination under the ownership of the merged entity of two datasets previously held by two independent firms. First, the combination of two datasets post-merger may increase the merged entity's market power in a hypothetical market for the supply of this data or increase barriers to entry/expansion in the market for actual or potential competitors, which may need this data to operate on this market. Competitors may indeed be required to collect a larger dataset in order to compete effectively with the merged entity than absent the merger. Second, even if there is no intention or technical possibility to combine the two datasets, it may be that pre-merger the two companies were competing with each other on the basis of the data they controlled and this competition would be eliminated by the merger." (European Commission 2016b)
Apple/Shazam	p. 7: "5.2.1.1. Access to commercially sensitive information (33) Shazam currently collects certain data on users of third party's apps, and in particular digital music streaming apps, installed on the same smart mobile devices where the Shazam app is installed. Through this data, post-Transaction, Apple could thus derive a list of customers of Apple Music's rivals on non-iOS devices, notably Android (1). [...]" (European Commission 2018b)
Microsoft/Github	p. 21: "5.4.3. Vertical non-coordinated effects regarding access to data 5.4.3.1. Potential concern (131) GitHub collects three categories of data: user-generated content, users' personal information, and metadata. [...]" (European Commission 2018a)
Google/Fitbit	pp. 97ff.: "9.3.3. Fitbit as source of data for possible use in online advertising services [...] 9.3.3.2.3. Effects of the data combination (444) As regards the effects of the data combination in the various markets for the supply of online advertising services [...]." (European Commission 2020)
Microsoft/Nuance	pp. 31f.: "5.2.3. Possible foreclosure effects in relation to Nuance's medical data (input foreclosure) [...] as set out in Apple/Shazam, the Commission notes that there are certain regulatory limitations to prevent the illegal combination of datasets." (European Commission 2021)
Meta/Kustomer	p. 100ff.: "(b) Incentive to foreclose [...] B. Additional data for online ads purposes (323) By engaging in a foreclosure strategy and steering businesses away from their current CRM provider to Kustomer, Meta (formerly Facebook) would also gain additional data from these business customers. [...]" (European Commission 2022)

TABLE 9 Top Data Firms

This table presents the top data firms for the fiscal year 2021 (most recent year in the sample) according to their normalized data intensity scores.

	CIK	Firm Name	Fama French 49 Industry	Data Intensity
1	1777921	AvePoint	Computer Software	0.993
2	1048695	F5	Computer Software	0.934
3	1544522	Freshworks	Computer Software	0.933
4	1373715	ServiceNow	Computer Software	0.919
5	1124610	VMware	Computer Software	0.910
6	1739942	SolarWinds	Computer Software	0.910
7	896878	Intuit	Computer Software	0.906
8	858877	Cisco Systems	Computer Hardware	0.902
9	1645590	Hewlett Packard Enterprise (HPE)	Computer Hardware	0.900
10	1447669	Twilio	Computer Software	0.895

TABLE 10 Data Intensity Percentile of Well-known Technology Companies

This table presents the percentiles of maximum data intensity scores of the 'big 5' and further U.S. firms in my sample listed in the Nasdaq Yewno Global Artificial Intelligence and Big Data Index (NYGBIG®) as of September 2023.

Firm Name	CIK	FF49 Industry	%ile Data Intensity
		'Big Five'	
Alphabet/Google	1652044	Computer Software	100
Apple	320193	Computer Hardware	99
Meta/Facebook	1326801	Computer Software	98
Amazon.com	1018724	Retail	93
Microsoft	789019	Computer Software	100
		Further U.S. firms listed in the (NYGBIG®)	
Adobe	796343	Computer Software	97
Advanced Micro Devices	2488	Electronic Equipment	93
Alarm.com	1459200	Computer Software	85
Ambarella	1280263	Electronic Equipment	83
Arista Networks	1596532	Computer Hardware	98
Asana	1477720	Computer Software	94
AT&T	732717	Communication	71
Bank of America	70858	Banking	59
Cadence Design Systems	813672	Computer Software	98
Ciena	936395	Electronic Equipment	98
Cisco Systems	858877	Computer Hardware	100
Commvault Systems	1169561	Computer Software	98
Crowdstrike	1535527	Computer Software	97
Dolby Labs	1308547	Trading	89
Dropbox	1467623	Computer Software	96
eBay	1065088	Computer Software	89
F5	1048695	Computer Software	100
Fastly	1517413	Computer Software	99
Fortinet	1262039	Computer Hardware	99
GenDigital/ Norton Life Lock	849399	Computer Software	100
Hewlett Packard Enterprise	1645590	Computer Hardware	100
Intel	50863	Electronic Equipment	99
IBM	51143	Computer Software	89
Intuit	896878	Computer Software	100
Juniper Networks	1043604	Computer Hardware	100
Micron Technology	723125	Electronic Equipment	89
MicroStrategy	1050446	Computer Software	96
Motorola Solutions	68505	Electronic Equipment	97
NetApp	1002047	Computer Hardware	100
Netscout Systems	1078075	Computer Software	100
Nutanix	1618732	Computer Software	93
Nvidia	1045810	Electronic Equipment	99
Oracle	1341439	Computer Software	94
Palantir	1321655	Computer Software	80
Palo Alto Networks	1327567	Computer Hardware	98
Pure Storage	1474432	Computer Hardware	100
Salesforce	1108524	Computer Software	99
ServiceNow	1373715	Computer Software	100
Silicon Labs	1038074	Electronic Equipment	92
Snap	1564408	Computer Software	97
Snowflake	1640147	Computer Software	88
Splunk	1353283	Computer Software	100
Synaptics	817720	Electronic Equipment	97
Synopsis	883241	Computer Software	99
Tenable	1660280	Computer Software	87
Teradata	816761	Computer Software	97
Uber Technologies	1543151	Transportation	65
UIPath	1734722	Computer Software	98
Verizon Communications	732712	Communication	85
Western Digital	106040	Computer Hardware	91

TABLE 11 Pairwise Correlations

This table presents the pairwise correlations between all variables presented in Table 1. Acquirer (0/1) turns to 1 if firm was an acquirer in a majority deal at time t+1; Target (0/1) turns to 1 if firm was a target in a majority deal at time t+1; Data Intensity refers to the data intensity measure presented in the data and methods section; R&D Intensity is the 3-year average SIC4-industry median adjusted R&D to sales (see King et al. 2008, building on Dierickx and Cool 1989; Cohen and Levinthal 1989, 1990); R&D Stock refers to the logarithmized 3-year depreciated (15% rate) sum of R&D investments (see King et al. 2008) while R&D is set to zero if missing; No R&D is a dummy equal to one if R&D was missing in data; Tobin's Q refers to the natural logarithm of the market to book value of firm assets; Tangible is property, plant, and equipment to total assets; Cash is cash to total assets; Leverage is market value of equity to market value of firm assets; ROA equals net income to total assets; Sales is the natural logarithm of firm sales as firm size indicator; HHI is the sales based Herfindahl Index within firm's TNIC-3 industry network; P/E Ratio represents market value of equity to net income; Age represents Compustat firm age, calculated from the observation year minus first Compustat listing year; Intangible represents other intangible assets (i.e., excl. goodwill, etc.) to total assets; Lifecycle (1/5) represents Dickinson (2011)'s life cycle measure attributes the stages *Introduction*, *Growth*, *Mature*, *Shake-Out*, and *Decline* to firm-fiscal year observations according to the signs of cash flows from operating, financing, and investing activities.

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. Acquirer (0/1)	1																
2. Target (0/1)	-0.07	1															
3. Data intensity	0.16	0.04	1														
4. R&D intensity	-0.07	-0.01	-0.01	1													
5. R&D stock	0.15	0.01	0.54	0.14	1												
6. No R&D (0/1)	-0.06	-0.02	-0.46	-0.13	-0.71	1											
7. Tobin's Q	0.08	-0.03	0.21	0.12	0.32	-0.29	1										
8. Tangible	-0.07	-0.01	-0.34	-0.07	-0.19	0.13	-0.09	1									
9. Cash	-0.07	0.01	0.24	0.37	0.36	-0.34	0.42	-0.28	1								
10. Leverage	-0.08	0.01	-0.35	-0.14	-0.39	0.44	-0.55	-0.03	-0.48	1							
11. ROA	0.14	-0.02	-0.04	-0.45	-0.11	0.14	0.07	0.06	-0.34	0.08	1						
12. Sales	0.25	-0.05	0.07	-0.32	0.12	0.04	-0.03	0.19	-0.40	0.12	0.43	1					
13. HHI	0.03	-0.03	0.16	-0.06	0.05	-0.15	0.05	0.02	-0.04	-0.20	0.06	-0.03	1				
14. PE ratio	0.05	0.00	0.02	-0.05	0.00	0.00	0.07	-0.01	-0.05	-0.03	0.13	0.08	0.01	1			
15. Age	0.07	-0.05	0.00	-0.10	0.10	-0.01	-0.03	0.16	-0.19	0.00	0.18	0.39	0.19	0.03	1		
16. Intangible	0.13	0.01	0.23	-0.04	0.17	-0.13	0.04	-0.15	-0.14	-0.11	0.00	0.14	0.10	0.01	0.03	1	
17. Lifecycle (1/5)	-0.03	0.01	0.00	-0.01	0.03	-0.01	-0.04	-0.05	0.06	-0.09	0.05	0.00	0.03	-0.01	0.05	-0.02	1

TABLE 12 Pairwise Correlations Data Intensity Measures

This table presents the pairwise correlations between the different data intensity variables. For measuring data intensity (1.), the cosine similarity to 10-Ks items 1 and 1A of the firms Alphabet/Google, Meta/Facebook, Microsoft, Apple, LinkedIn, Fitbit, and Nuance is measured. For the NYGBIG® measures, benchmark firms are all U.S. listed firms in the Nasdaq Yewno Global AI and Big Data index. The intensity measure (2.) includes all 55 U.S. firms included in the NYGBIG® measures as of September 2023. These 55 firms are additionally subdivided in B2C firms (3.), i.e., companies with consumer interaction and thus potentially collecting primary customer/user data, and B2B firms (4.) without direct consumer interaction.

Variable	Mean	S.D.	1.	2.	3.	4.
1. Data Intensity	0.390	0.177	1			
2. NYGBIG® Intensity	0.403	0.182	0.977	1		
3. NYGBIG® B2C	0.426	0.173	0.959	0.971	1	
4. NYGBIG® B2B	0.389	0.184	0.968	0.996	0.944	1