Breaking Network Barriers in the Era of Data-Driven Venture Capitalists

Melissa Crumling*

Drexel University

November 27, 2024

Abstract

Information frictions in financial decision-making are particularly salient in the venture capital (VC) industry, where VCs traditionally rely on professional networks to identify potential investment opportunities. Over the past decade, however, VCs have increasingly adopted digital data and machine learning techniques to inform their investment decisions, marking a significant shift from traditional methods. I posit that these technologies, capable of identifying all startups with a digital presence, reduce information frictions in identifying promising ventures. Using the geographic concentration of the VC industry as my empirical setting, I find that VCs are more likely to invest outside traditional VC hubs after adopting data technologies. Moreover, these investments are more likely to exit through an IPO or achieve unicorn status than their counterparts in established hubs. Additionally, data technologies help locate startups in areas outside major hubs with increasing entrepreneurial activity, which subsequently experience growth in VC activity. These findings highlight the benefits of using data technologies to identify promising ventures, benefiting both investors and emerging VC markets.

JEL Classification: D85, G24, G32, L26, O32, O33

Keywords: Entrepreneurial Finance, Venture Capital, Networks, Big Data, Innovation

*PhD Candidate, Drexel University, LeBow College of Business, email: mej69@drexel.edu. I thank my advisors Michelle Lowry (chair), David Becher, Daniel Dorn, Michael Ewens, and Greg Nini for their guidance and support. I also thank Natee Amornsiripanitch (Philadelphia Fed mentor), Jason (Pang-Li) Chen, Winston Dou (Rising Scholars mentor), Casey Dougal (discussant), Thorsten Drautzburg, Tanja Kirmse, Alice Liu (discussant), Torin McFarland, Adriano Rampini, Chris Severen, Alex Tuft (discussant) and participants of the Rising Scholars Conference 2023, WEFI (Workshop for Entrepreneurial Finance and Innovation) Fellows 2024, Philadelphia Fed Mentoring and Advancing PhDs program 2024, and Financial Management Association (Doctoral Consortium) 2024 for helpful comments.

1 Introduction

Managing a portfolio of assets involves acquiring information and using that information to inform investment decisions. However, information acquisition is costly (Grossman and Stiglitz (1980), Verrecchia (1982)), prompting a large literature to explore mechanisms to reduce these information barriers. For instance, geographic proximity (e.g. Huberman (2001)) and network centrality (e.g. Kuchler et al. (2022)) have been shown to facilitate investment opportunities by improving access to information. Recently, the rise of data accessibility and advancements in data technologies have radically changed information flow. Investors can now access company data within seconds, and tools such as machine learning and artificial intelligence can synthesize large amounts of data to inform investment decisions. This raises the following research questions: do data technologies reduce information frictions for financial decision-making, and how does this impact who receives funding?

To explore this question, I examine the role of data technologies in reducing information asymmetry within the venture capital (VC) industry. I study the VC market due to the importance of the setting, the centrality of deal flow in the investment process, and the presence of an ideal empirical setting. First, VCs are crucial providers of capital to young, innovative firms, with approximately 50% of all publicly traded companies having received VC financing prior to the IPO (Gornall and Strebulaev (2021)). Second, information frictions are particularly salient in the VC industry, as VCs invest in young companies with limited track records. Lastly, the VC industry is highly geographically concentrated, with over two-thirds of VC activity centered in three main areas: California, Massachusetts, and New York (NVCA (2020)). This provides an ideal empirical setting to test the impact of data technologies on overcoming search frictions.

To identify if and when VCs adopt data technologies, I utilize detailed employee data from Crunchbase and LinkedIn. VCs using data technologies rely on human capital and expertise to implement the data infrastructure. Prior research has used job postings to infer technology adoption in other settings[1] and specifically in the VC domain (Retterath (2020), Bonelli (2023)).

[1] For example, job posting data has been used to identify demand for AI skilled labor in public firms (Alekseeva et al. (2021), Babina et al. (2024)) and in the real estate industry (Raymond (2024)) and Goldfarb, Taska, and

Using job titles and descriptions from a complete history of VC employees, I identify when VCs hire data scientists and classify a VC firm as data-driven from the date of its first data-related employee hire.[2] Using data technologies in the pre-investment screening process represents a significant shift from traditional deal-sourcing methods, which historically rely mostly on professional networks (Gompers et al. (2020)). As shown in Figure 1), VCs are increasingly adopting data-driven approaches to aid their investment decision processes. Industry reports predict 75% of VCs will use data technologies in some capacity by 2025 (Gartner (2023)).

In the first part of the paper, I investigate whether data technologies impact VCs' investment opportunity set. My overarching prediction is that adopting data technologies lowers search frictions for finding investment opportunities, as VCs are able to find all potential investments with an online presence. I use startups located outside of traditional VC hubs (California, Massachusetts, and New York) as a proxy for those outside of established VC networks. VC activity is highly concentrated in traditional hubs, with 85% of capital raised by VCs and 73% of capital invested in startups located in these areas (NVCA, 2020). VCs tend to invest locally (Sorenson and Stuart 2001) as geographic barriers facilitate information flow about potential investment opportunities (Cumming and Dai 2010). Additionally, VCs are more likely to establish satellite offices in these regions (Chen et al. 2010), further enforcing hub-based interactions across geographic markets. As a result, startups outside these areas are more likely to be excluded from traditional networks and face larger information frictions. My findings support this prediction: after VCs adopt data technologies, they increase their number of investments in non-hub locations by 15% per year. This translates to an increase of approximately half an investment per year, a non-trivial amount considering the industry's high geographic concentration. In additional analyses, I further narrow non-hub locations to those with fewer than 25 VC investments over the past five years (classified as low-activity areas). After adopting data technologies, VCs increase their investments in these low-activity areas by 60%, or approximately 0.15 investments per year. These findings provide evidence that data technologies increase VCs potential investment opportunity set to startups

Teodoridis (2021))

[2] Alternatively, VCs could hire data scientists that use AI to help their startup companies —a classification I am careful to exclude.

that would otherwise be excluded from their professional networks.

In my second set of tests, I examine other proxies for startups that would fall outside of a VCs professional network. I test whether VCs rely less on other investors to find investment opportunities after adopting data technologies. VCs tend to syndicate investments with other investors, a practice that helps overcome information frictions (Lerner (2022)). Conditional on investing in a different state, I find that VCs are 4-7% less likely to invest with a local VC syndicate. VC networks can also vary at the industry level. A large literature shows that VCs tend to specialize in investing in certain industries (e.g. Hochberg, Mazzeo, and McDevitt (2015)) and these industries can form established networks (Hochberg, Ljungqvist, and Lu (2010)). I classify VCs as specializing in a particular industry if more than 40% of their investments were in one industry over the last five years. I find that after adopting data technologies, VCs that specialized in one industry are approximately 40% more likely to invest in a different industry.

I implement various strategies to mitigate concerns that the results are driven by correlated unobservables. To address concerns that VCs using data technologies may differ from those that do not, I include VC firm fixed effects, allowing for a comparison of VC investment decisions before and after technology adoption. Additionally, I include VC headquarter state × industry × funding round stage × investment year fixed effects to account for local time trends coinciding with VCs' adoption of data technologies that may lead them to invest in non-hub locations. To mitigate concerns that data technology adoption is correlated with overall firm growth, I conduct a placebo test in which I repeat the same analysis but use the hiring of a venture partner in place of a data scientist. The hiring of a venture partner is associated with an increase in the number of investments, but only in hub areas, not in non-hub or low-activity commuting zones. This provides further evidence that data technologies help overcome information frictions in identifying investment opportunities compared to an increase in the VCs' human capital, which leads to more hub investments. Lastly, to address concerns about selection bias, I employ an instrumental variables approach. Specifically, I isolate variation in VCs' data technology adoption from two sources: early exposure to AI and the timing of raising a new fund. This identification strategy reduces bias from demand shocks that could influence both technology adoption and investment strategies. The results remain robust to the IV approach, providing evidence of a causal relationship between data technology adoption and investment decisions.

A natural follow-up question is: how do data-driven investments in non-hub areas perform compared to hub investments? Prior literature indicates that data technology adoption provides an advantage in identifying startups less likely to fail but does not confer any additional advantage in finding startups more likely to achieve a major exit, such as an IPO or acquisition (Bonelli (2023)). However, little is known about how data-driven investments in non-hub and low-activity areas compare to (1) their data-driven hub counterparts and (2) traditional investments in non-hub and low-activity areas. Ex ante, the performance outcomes of data-driven non-hub investments are ambiguous. On one hand, VCs investing in non-hubs may face less competition, potentially enabling them to invest in higher-quality startups. This aligns with prior literature showing that VCs set a higher hurdle rate for non-hub startups, which tend to outperform their hub investments (Chen et al. (2010)). Alternatively, if data technologies reduce search frictions in finding investments, they may allow VCs to identify more firms in non-hub areas, including more lower-quality startups. Although VCs actively monitor their portfolio companies-a practice associated with increased performance (Bernstein, Giroud, and Townsend (2016))-effective monitoring may be more challenging with distant, non-hub investments, especially for lower-quality firms.

I find evidence supporting both scenarios: data-driven investments in non-hub areas are more likely to fail, but they are also more likely to achieve a major success. The results for the failure analysis offer two interpretations. First, data-driven VCs may invest in lower-quality non-hub startups compared to their hub investments. Alternatively, data-driven VCs might abandon their lower quality non-hub startups more readily than those in hub locations. Local investments allow VCs to monitor more effectively (Lerner (1994)), potentially offsetting concerns about quality. Either way, VCs generate the majority of their returns through a few "home-run" investments, and I find evidence that data technologies provide advantages in identifying these types of startups in non-hub locations.

In the final part of the paper, I investigate which non hub areas are likely to attract data-driven investments and if data-driven investments in these areas lead to an increase in subsequent VC activity. One advantage of algorithmic techniques is that they are able to identify emerging trends and markets. Thus I posit that data technologies are more likely to identify promising investments in areas with growing levels of entrepreneurial activity. Using regional entrepreneurial statistics from the Startup Cartography Project (Fazio et al. (2019)), I find that non-hub areas with growing levels of entrepreneurial activity receive more data-driven investments than areas with low levels of entrepreneurial activity. In contrast, I find no evidence that traditional VC investments can identify growing areas of entrepreneurial activity. In addition, I find that low-activity areas experience an increase in subsequent VC activity after a startup in that area receives a data-driven investment. Specifically, I find these low-activity areas experience an increase in the number of funding rounds, the number of startups that receive VC financing, the number of investors and the number of VC-backed firm patents after entry by a data-driven VC compared to low activity areas that do not receive a data-driven investment. In sum, my results indicate that data-technologies are better able to identify emerging markets and trends than traditional methods, and that these areas experience an increase in subsequent VC activity.

The rest of the paper proceeds as follows. Section 2 discusses my findings' contribution to relevant literature. Section 3 discusses the institutional background. Section 4 details data sources and construction of measures. Section 5 reports my main findings on how VCs' investments change after adopting data technologies. Section 6 investigates how data-driven investments perform in non hub locations. Section 7 identifies which non-hub areas receive data-driven investments. Section 8 concludes.

2 Contribution to Prior Literature

This paper relates to several strands of literature. First, this paper contributes to the literature studying mechanisms to reduce information frictions in financial markets. Even in public markets, where disclosure is enforced, investors tend to display a large home bias in investments (Van Nieuwerburgh and Veldkamp (2009)), which can lead to information advantages in stock selection (Coval and Moskowitz (2001)) but also underdiversification in investor portfolios (Huberman (2001), Van Nieuwerburgh and Veldkamp (2010)). Social networks also play an important role in information transmission in financial markets (e.g Hong and Xu (2019)). However, for both geographic and social networks, the literature provides conflicting evidence on whether these networks provide superior information for returns (Massa and Simonov (2006), Cohen, Frazzini, and Malloy (2008), Seasholes and Zhu (2010), Pool, Stoffman, and Yonker (2012), Pool, Stoffman, and Yonker (2015), Kuchler et al. (2022)). In recent years, the role of technological advancements has also impacted information frictions in financial markets. Introduction of commercial databases reduces barriers for information acquisition (Gao and Huang (2020)), and more recently the rise of AI has helped inform investment decisions (Cao et al. (2024)). This paper contributes to the literature by studying the role of advanced technological techniques for finding investment opportunities in an industry with limited disclosure, specifically the VC industry.

Second, this paper adds to the literature on how VCs source investments. Considered one of the most important factors of deal success (Sørensen (2007)), 60% of investments come from a VCs' network (Gompers et al. (2020)). Strong networks between VCs allow for better fund performance (Hochberg, Ljungqvist, and Lu (2007)) and can create extensive barriers to entry for new VCs firms in existing markets (Hochberg, Ljungqvist, and Lu (2010)). However they can also be used to overcome geographic barriers through syndicated investments (Sorenson and Stuart (2001)) or alumni networks with founders (Garfinkel et al. (2021), Huang (2022)). The consequences of strong networks is that the capital for innovation is largely centralized in a few distinct locations in the US (Lerner and Nanda (2020)) which can impact the innovation prospect of other economies (Glaeser, Kerr, and Ponzetto (2010)). This paper studies the implications of adopting data technologies as another means to overcome information frictions when sourcing investments.

Third, this paper also contributes to the literature studying the role of data technologies in the VC industry. Prior literature has looked at the role of the internet (e.g. Li, Li, and Yang (2022)) and direct airline routes (Bernstein, Giroud, and Townsend (2016)) on finding investment opportunities. Recent literature looks at the role of artificial intelligence in the VC industry. Lyonnet and Stern (2022) and Davenport (2022) look ex ante how algorithms could be used to outperform human investments in startups. They use machine learning to identify the most promising ventures and find that VCs invest in some firms that perform predictably poorly and pass on others that perform predictably well largely due to stereotypical thinking by VCs. Retterath (2020) develops an algorithm to predict successful investments in the VC industry which outperforms that of actual investments. The only other paper (to my knowledge) that looks at the ex post impact of data technologies on investment decisions is Bonelli (2023), who finds that VCs are more likely to invest in startups similar to their previous investments and less in break through technologies. While this study evaluates the screening ability of data technologies, I look at how data technologies lower search costs and the overall impact this has on the financing of innovation.

Lastly, this paper contributes to the growing of data technologies in financial markets. Prior research has examined these technologies in the banking sector and credit markets (Fuster et al. (2022); Blattner and Nelson (2021); Di Maggio, Ratnadiwakara, and Carmichael (2022)), financial analysts (Birru, Gokkaya, and Liu (2018); Coleman, Merkley, and Pacelli (2021); Grennan and Michaely (2020); Dessaint, Foucault, and Frésard (2021); Chi, Hwang, and Zheng (2023)), asset management (DâAcunto, Prabhala, and Rossi (2019); Rossi and Utkus (2020); Abis (2020); Abis and Veldkamp (2024)) and stock price information dissemination (Bai, Philippon, and Savov (2016); Dugast and Foucault (2018); Zhu (2019); Farboodi and Veldkamp 2020; Gao and Huang (2020); Farboodi et al. (2022)). This paper investigates the impact of data technologies in the VC industry.

3 Institutional Background - Traditional VC Model vs Data Driven VCs

VC activities encompass three primary tasks as outlined by Gompers et al. (2020): (i) preliminary investment screening, which involves sourcing, evaluating, and selecting investments, (ii) investment structuring, and (iii) post-investment value enhancement, including activities like monitoring and advising startups. Traditionally, pre-investment screening, which plays the most crucial role in value creation (Sørensen (2007); Gompers et al. (2020)), relies heavily on existing networks (Hochberg, Ljungqvist, and Lu (2007); Howell and Nanda (2019)) and subjective assessments by VC partners (Kaplan and Strömberg (2000); Kaplan, Sensoy, and Strömberg (2009); Lyonnet and Stern (2022); Gompers et al. 2022). However, evaluating hundreds of startups annually can be lengthy and time-consuming. Many firms therefore adopt data technologies to automate parts of the pre-investment screening process.

Specifically for sourcing deal flow, VCs want to identify as many startups as possible, to maximize the likelihood of finding a "home run" investment. However in practice, VCs largely rely on inbound approaches to find investment opportunities. Gompers et al. (2020) find that approximately 60% of deals are sourced through a VCs' professional network. As Damian Cristian, founder of Koble — a sourcing platform engineered by AI — puts it "*Above all else, VC is a network business, effectively capped by the scalability of human relationships*". While better networked VCs are shown to have superior fund performance (Hochberg, Ljungqvist, and Lu (2007)), the coverage of potential investments is largely incomplete which can prevent best possible fit between startups and VCs. However, using data technologies allows the VC to find every company possible. Early adopters of these technologies hire data scientists to build their own internal data infrastructures. [3] Data scientists use data technologies to identify firms at their earliest stages and use web crawling tools to expand their search from commercial databases (such as Pitchbook and Crunchbase) to non-obvious sources like LinkedIn, Github, X, and new firm registrations. SignalFire's platform, BeaconAI, tracks and ranks more than 80 million companies, 600 million people and millions of open-source projects.[4]. Basis Set Ventures uses large language models to study founders' cognitive and behavioral traits to predict founder success.[5]. Data-driven approaches provide a competitive advantage to VCs over traditional methods, as they are no longer bound by their networks and can identify promising founders and markets more efficiently. Data technologies then use all the gathered information to score the startups and provide informative metrics for VCs to decide which companies to invest in. While this paper focuses mostly on sourcing startups, see Bonelli

[3] For example, Signal Fire built Beacon AI, Tribe Capital built Termina, and Connetic Ventures built Wendel.

[4] https://www.signalfire.com/blog/signalfire-beacon-ai

[5] https://www.basisset.com/founder-superpowers

(2023) for more information on how data technologies are used in the screening portion of the pre-investment screening of startups. To end the quote by Damian Cristian "*There is a cognitive gap in how many sectors and companies an individual investor can deeply understand without the help of data and technology. Technology solves this limitation, enabling investors to source and screen huge deal flow volumes.*"

4 Data and Summary Statistics

4.1 VC Investments

I use data from Crunchbase to construct my investment sample. Crunchbase is an online database providing detailed information on startup firms and their investors. I start by defining my VC investor sample. I keep all VC firms headquartered in the US and defined as venture capitalists, micro venture capitalists or private equity firms [6]. I then merge the remaining VCs with Preqin and VentureXpert to ensure coverage in multiple databases. I am left with 927 distinct VC firms during my sample period of 2010 to 2022 [7]. For each VC firm, I gather information on their founding year, headquarter location, assets under management, and full employee and job histories provided in Crunchbase. After identifying my sample of investors, I use all their investments made in the US after 2010. I restrict my sample of investments to those classified as pre-seed, seed, and series a, b, c, and d+. My final sample amounts to 927 unique investors, 8,513 VC-years, and 62,020 VC investments.

Lastly, I gather information on all the startups invested in by my VC sample. This includes their founding year, industry classification, head quarter location and founder information from their employee and job histories. Founder information includes gender, education, and whether they are a serial entrepreneur or previously a VC. I also follow methodology on an emerging literature on VC investment and alumni networks to identify whether VC partners and startup founders attended the same alma mater (Garfinkel et al. (2021), Huang (2022), and

[6] I exclude all firms classified as angel groups, family offices, funds of funds, investment banks, hedge funds, accelerators and incubators, government offices, university and entrepreneurship programs, coworking spaces, startup competitions, pension funds and loyalty programs.

[7] Crunchbase's coverage of startups has been validated to be most accurate in more recent years (Wu, 2016; Ferrati and Muffatto, 2020).

Koenig (2022)). My final sample includes 29,375 distinct startups that were at some point VC funded.

4.2 Methodology to Identify Data-Driven VCs

Following prior literature (e.g. Bonelli (2023), Retterath (2020), Raymond (2024)), I define VC firms utilizing data technologies as those that hire data scientists or data-related employees. The rationale is that VCs leveraging data technologies depend on human capital and expertise to implement and maintain their data infrastructure. I employ a three-step process to identify when VCs become data-driven.

In step one, I compile a list of data-driven VCs from the "Data-Driven VC" website[8], an initiative led by Andre Retterath, PhD, from Earlybird Ventures. This site provides insights for VCs interested in adopting data-driven approaches and publishes a weekly newsletter and an annual report on data-driven practices in the VC industry. According to their methodology, a VC is classified as data-driven if it meets three criteria: (1) employs at least one data engineer, (2) receives at least one community nomination as data-driven, and (3) has developed internal tooling across one or more segments of the VC value chain (e.g. sourcing, screening, due diligence, portfolio management, or exits). As of the 2024 Annual Report, 79% of identified data-driven VCs adopt data strategies to improve deal coverage[9]. Additionally, a 2018 Pitchbook survey found that 85% of VCs use data for sourcing investments, with 38% using data for all investment sourcing[10]. This indicates that data-driven approaches are primarily employed during the pre-investment screening process.

According to the Data-Driven VC criteria, 75 U.S.-based VC firms are classified as data-driven (out of 183 globally). I merge this list with my Crunchbase sample, resulting in 40 matches. The remainder of the firms were either classified as data-driven after my sample period (i.e., in 2023 or 2024) or excluded due to differences in classification, such as incubators. This refined list of data-driven VCs allows me to proceed to step two.

In step two, I scrape LinkedIn profiles to gather data on current and past employees of the

[8] https://www.datadrivenvc.io/

[9] https://landscape2024.datadrivenvc.io/

[10] https://pitchbook.com/media/press-releases/pitchbook-survey-finds-only-38-of-venture-capital-investors-currently-use-da

identified data-driven VCs. I manually review each employee profile to identify those working as data scientists or in closely related roles, ensuring they were hired to develop or manage internal tools for pre-investment screening . I also compile a list of relevant job titles from these data-related employees.

In step three, I use Crunchbase and LinkedIn to identify profiles for all current and past employees of the remaining VCs in my sample. Using the job titles identified in step two, I flag any data-related employees and manually verify whether they meet the internal tooling criteria set by Data-Driven VC (i.e. hired to build or manage data-related internal tools used in pre-investment screening). This process identifies an additional 13 data-driven VCs. Following prior literature, I classify a VC as data-driven based on the hire date of its first data-related employee, resulting in 59 data-driven VCs. All other VCs are considered traditional. This classification results in 2,964 data-driven investments. Panel A of Table 2 shows the mean difference between data-driven and traditional (i.e. non data-driven) VCs in my sample. Data-driven VCs are considerably older (14.65) than traditional VCs (11.92) and are also larger based on number of employees (23.18 versus 8.95) and assets under management (1.3 billion versus 0.4 billion). They also tend to have larger networks (5.93 versus 2.69) and invest in more startups per year (22.47 versus 7.85).

5 Data-Driven Investors and Investment Opportunities

In this section, I examine whether data technologies reduce search frictions in identifying investment opportunities. VCs invest in startups, which by nature, are characterized by high levels of information asymmetry Dessein (2005), This leads VCs to favor local investments Chen et al. (2010), as geographic proximity facilitates information flow. Since data technologies can identify all startups with a digital presence, this section explores whether these technologies help overcome traditional geographic barriers in VC investing.

5.1 Quantity of Investments

First, without claiming causality, I investigate whether data technologies scale investors' opportunities set. Without access to deal flow data, I proxy for this by looking at the VC's overall investments in a given year as well as where they choose to invest. The intuition is that data technologies are able to find all possible startups with an online presence and can therefore identify potential investments that would fall outside a VCs network. The VC industry is highly concentrated with over 79% of capital invested in California, New York, and Massachusetts (Lerner (2010)). I therefore define VC "Hubs" as commuting zones located in San Francisco and San Jose, California; New York, New York, and Boston Massachusetts. I define "Non Hubs" as commuting zones other than Hub commuting zones and "Low Activity" commuting zones as those with 25 or fewer VC investments over the previous 5 years. I use 25 or fewer investments as prior literature has used this cutoff to define established VC markets (e.g. Hochberg, Ljungqvist, and Lu (2010)). I estimate the following regression at the VC-Year level:

$$Y_{j,t} = \beta DataDriven_{j,t} + X_{j,t} + \alpha_j + \alpha_{c \times i \times s \times t} + \epsilon_{j,t} \tag{1}$$

The dependent variable is the number of investments made by VC j in year t. The main explanatory variable, $DataDriven$, is a dummy variable equal to 1 if VC j is classified as data-driven as of year t and 0 otherwise. $X_{j,t}$ are time varying controls of VC j, including the VC-firm age (controlling for experience), the number of employees and total assets under management (controlling for size), and their eigenvector centrality (controlling for network intensity). α_j are VC firm fixed effects to control for any time invariant VC characteristics. $\alpha_{c \times i \times s \times t}$ are VC-headquarter state c × startup industry i × funding stage s × funding year t fixed effects to alleviate concerns of the VC's location, time and industry trends coinciding with VC's adoption of data technologies that leads them to invest in startups outside major hubs. The coefficient β is therefore estimated by comparing VC j's investments before versus after data technology adoption relative to other VC firms' investments in the same state-industry-stage-year segment. Since the number of investments is a count variable with certain specifications left-censored at zero and skewed, I estimate a Poisson model. Standard errors are clustered at the VC-firm level.

I begin by looking at the overall number of investments made by a VC in a given year. Table 3 Panel A reports the results. Since some VCs are founded as quantitative firms (i.e. have always been data-driven), the first four columns do not include the VC-firm fixed effect to conduct a between firm analysis. Column (1) estimates Equation 1 without controls. The coefficient on $DataDriven$ is positive and significant with a sizable magnitude, indicating that data-driven VCs make $e^{0.790} - 1 = 120\%$ more investments than traditional VCs. However, the size and significance of this coefficient is largely dependent on time-varying VC controls. In column (2), I include the natural log of VC firm age to proxy for experience. The coefficient is positive and statistically significant and decreases the coefficient on $DataDriven$ to 0.688. In column (3), I include the natural log of the number of employees at the VC firm and the natural log of the total assets under management to proxy for firm size. Both coefficients are positive and statistically significant, indicating that larger VC firms make more investments. The coefficient on $DataDriven$ decreases substantially to 0.113 and is no longer statistically significant. The coefficient on $Log(VCFirmAge)$ also decreases in magnitude, switching signs to negative and becomes statistically insignificant. In column (4), I include the VC firm's eigenvector centrality to proxy for the size of their network. Prior literature finds that VCs' investments are largely sourced through inbound approaches, and the coefficient on $Centrality$ is positive and significant, indicating VCs with larger networks invest more. The coefficient on $DataDriven$ falls to 0.026 and is largely insignificant with t-statistic of 0.46. Thus firm size and network centrality mostly explains the differences in the number of investments made by a VC in a given year.

In columns (5) through (8) I include VC-firm fixed effects. This allows me to compare the number of investments VCs make before and after they adopt data-technologies. The coefficients across the specifications are more stable, even after including proxies for VC firm age, size, and network centrality. The most stringent specification is column (8) and the coefficient on data-driven can be interpreted as, after VCs adopt data technologies, they increase the number of investments made by approximately 14%, compared to VCs in the same state-industry-stage-year segment. The average VC therefore increases the number of investments from 8 to 9 investments per year. The persistent significance and large magnitudes on proxies for VC firm size and network centrality offer important inferences for VC-investment decisions. Overall, the results suggest that data-driven VCs do not invest in more startups relative to traditional VCs, however, after they adopt data technologies, they increase their number of investments on average by an additional firm per year.

In Panel B of Table 3, I replace the dependent variable with the number of investments in hub commuting zones (i.e., San Francisco and San Jose, CA; Boston, MA; and New York, NY). Before controlling for firm size and network centrality (columns (1), (2), (5), and (6)), the coefficients are positive and statistically significant, indicating that data-driven firms increase their investments in hub locations after adopting data technologies. VCs tend to concentrate the majority of their investments in these areas, with an average of 5 out of 8 investments per year, and two-thirds of VC-backed startups are located in these regions (NVCA, 2020). Therefore, if VCs scale up their investments, it is intuitive that they would do so by investing in hub-located startups. However, the magnitudes decrease substantially and lose statistical significance after controlling for firm size and network centrality proxies. This suggests that the observed increases in hub investments are largely explained by a firm's network, rather than the adoption of data technologies. In contrast, Panel C presents the results when the dependent variable is replaced with the number of investments in non-hub commuting zones. The coefficients on $DataDriven$ in columns (1), (2), (5), and (6) are positive and statistically significant, though smaller in magnitude than those in Panel B. However, when size and centrality controls are included in columns (7) and (8), the coefficients remain consistent in both magnitude and significance. This indicates that the increase in investments in non-hub areas is not fully explained by a VC firm's size and network but is also attributed to the adoption of data technologies. The results are even stronger in Panel D, where I further restrict the number of investments to those made in low-activity areas. The most stringent specification, shown in column (8), suggests that after VCs adopt data technologies, they increase the number of investments in low-activity areas by $e^{0.46} - 1 = 58\%$.

To mitigate concerns that VCs located in non hub areas are driving results, I repeat the above analysis but exclude all VCs located in non bub areas. The results are displayed in Panel

A of Table A1. Results are largely consistent: after VCs adopt data technologies, VCs increase the number investments they make, specifically in non hub and low activity areas. Another concern is that the analysis in Table 3 includes follow-on investments, which are unlikely to be sourced through quantitative means. In Panel B of Table A1, I only include first time investments by VCs in startups. The results hold. In addition, results are robust to variations of the main independent $DataDriven$ variable. In panel A of Table A2, results are robust to a continuous measure of "Data-driven intensity", defined as the logarithm of one plus the number of data-related jobs at the VC firm. In panel B of Table A2, results continue to hold when using the number of data-related employees scaled by the number of partners at the VC firm. Overall, these results suggest that after adopting data technologies, VCs increase their investments, particularly in non-hub and low-activity areas, providing evidence that data technologies can lower search frictions in identifying investment opportunities.

5.1.1 Confounding Variable: VC Firm Growth

The previous results indicate that after adopting data technologies, VCs tend to increase their investments, particularly in non-hub and low-activity areas. A natural follow-up question is: how do they achieve this? VCs are constrained by both how much they can invest (i.e., fund size) and the number of companies they can actively manage (i.e., number of partners). To explore this further, I begin by examining the first limitation: fund size. VCs can either raise larger funds at the time of data technology adoption, allowing them to invest more. Alternatively, VCs could participate in smaller round sizes, thus increasing the number of startups they can invest in. Data-driven approaches allow VCs to evaluate more startups than human capital alone, potentially leading them to adopt "spray and pray" strategies, where they invest in a larger number of startups but reduce the amount of capital per investment to hedge against failure (Ewens, Nanda, and Rhodes-Kropf (2018)). I conduct an OLS regression using Equation 1, replacing the outcome variables with $Log(TotalAUM)$ and $Log(MedianRound\$)$. To better capture firm growth, I remove controls for firm size (i.e. $Log(\#Employees)$). The results are displayed in Panel A of Table 4. Odd-numbered columns exclude VC-firm fixed effects to capture between firm variation and even-numbered columns include VC-firm fixed effects to capture within firm variation. In columns (1) and (2), the coefficient on $DataDriven$ is positive and statistically significant, indicating that after VCs adopt data technologies, their AUM increases. In columns (3) and (4), I replace the dependent variable with the natural log of median round size. The coefficients on $DataDriven$ are largely insignificant. These results suggest that VCs are able to invest more because they have more committed capital, not because they are decreasing round sizes.

Another constraint on VCs' ability to invest more is their human capital. VCs actively manage their portfolios (Hsu (2004), Bernstein, Giroud, and Townsend (2016), often taking board seats and meeting regularly with portfolio companies. Therefore, if VCs are making more investments, they would likely need to hire more partners to manage the increased workload. To examine this, I replace the dependent variable in Equation 1 with $Log(\#Partners)$. The results are displayed in columns (1) and (2) in Panel B of Table 4. In both specifications, the results are positive but the magnitude and significance weakens when including the VC firm fixed effect. In columns (3) and (4), I replace the dependent variable with the number of investments per partner and perform a Poisson model. The results are insignificant. Overall, these findings suggest that VCs tend to hire more partners after adopting data technologies, consistent with the increase in committed capital, which enables them to make more investments.

The results above suggest that the timing of data technology adoption is correlated with overall firm growth. A potential concern is that VCs may be investing more, particularly in non-hub and low-activity areas, as a byproduct of firm growth rather than due to data technology adoption through the hiring of a data-related employee. While the previous section controls for firm size by including $Log(\#Employees)$ and $Log(TotalAUM)$, I conduct two additional analyses to further address this concern to provide added confidence in my findings.

Examining Pre-Trends First, I address concerns that VCs hire data scientists during a period of growth by examining pre-trends. I begin by performing a more formal stacked difference-in-differences (DiD) approach. Given that the hiring of data scientists is staggered, I estimate a stacked DiD specification (Gormley and Matsa (2011) and Baker, Larcker, and

Wang (2022)) using an unbalanced VC panel.. The event window spans 5 years before and after the hiring event (the panel is unbalanced as some VCs were founded less than 4 years before the event), and the analysis is conducted at the VC-year level.:

$$Y_{j,d,t} = \beta Treated_{j,d} \times Post_{d,t} + X_{j,d,t} + \alpha_{j \times d} + \alpha_{d \times c \times i \times s \times t} + \epsilon_{j,d,t} \tag{2}$$

$Y_{j,d,t}$ is the number of investments made by VC j in year t. $Treated$ is an indicator equal to one if VC j hires a data scientist during my sample and zero otherwise. $Post$ is an indicator that equals one post-hiring and zero otherwise. Controls and fixed effects remain the same as before. Results are displayed in columns (1) through (3) in Table 5 and are consistent using the stacked DiD design: after VCs adopt data technologies, they invest in more startups (Panel A), specifically in non hub (Panel C) and low activity (Panel D) areas. Similar as before, coefficient magnitudes are sensitive to the inclusion of VC network centrality (column (2)) and VC firm size (column (3)), indicating that these are important controls when investigating where VCs choose to invest. To examine more closely whether data scientists were hired simply to manage overall investment growth, I plot the coefficient dynamics in Figure 2. The figure is constructed by replacing $Post_{d,t}$ in Equation 2 with event time dummies:

$$Y_{j,d,t} = \sum_{k=-4,k\neq-1}^{5} \beta_k [Treated_{dc} \times Year(k)_c] + \alpha_{d,c} + X_{j,d,t} + \alpha_{j \times d} + \alpha_{d \times c \times i \times s \times t} + \epsilon_{j,d,t} \tag{3}$$

where $Year(k)_c$ indexes years since cohort d's hiring of a data scientist. Panel A of Figure 2 shows the results for all investments made by VC j in year t. The figure shows slight pre-trends for total investments (Panel A). Thus to mitigate concerns that overall firm growth is correlated both with the hiring of a data scientist and investments in non-hub areas, I conduct the following placebo analysis.

Placebo Analysis For the placebo analysis, I create a matched sample with another type of employee hire: venture partner. Venture partners are different from general partners, as they are specifically hired to help with deal sourcing and strategic guidance (Rezaei (2024)).

They do not act as general partners, who are more senior and committed employees and are involved in all aspects of venture capital financing. In addition, they are not involved in the final investment decision, making them an ideal equivalent to data scientists as they are primarily hired to help with the front end of investment process.

For each data-driven firm in my sample, I propensity score match to another VC firm that hires a venture partner in the same year. The matched-VC has to be located in the same state and invest in the same industry and stage as the data-driven VC. I propensity score match on the controls: VC firm age, number of employees, total assets under management, and network centrality, the year prior to employee hire. I conduct a one-to-one match with replacement, and results are robust to a 1-3 match in the appendix (Table A3). The summary statistics and mean differences to the data-driven sample are in Panel B of Table 2. There are no statistical differences between investor age, assets under management, and network centrality, however, data driven VCs are larger in terms of employee size than the matched sample. This can be attributed to the largest VC firms being classified as data-driven (e.g. Andreesson Horowitz, Norwest Venture Partners, and Sequoia). In untabulated results, the median differences are not statistically significant from zero. I conduct the same analysis in Equation 2 as well as the dynamic analysis in Equation 3, but $Treated_{dc}$ is an indicator to a matched firm hiring a venture partner.

Panel A of Figure 3 shows the dynamics for total investments after the hiring of a venture partner. Similar to the data-driven analysis, the hiring of a venture partner is associated with an overall increase in the number of investments with a slight pre-trend. This indicates that hiring of a venture partner is also associated with overall firm growth. The coefficients are displayed in columns (4) through (6) of Panel A in Table 5. The coefficients are positive and statistically significant, indicating that after VCs hire a venture partner, they increase the number of investments by 11% to 17%. Similar to the data-driven results, the magnitudes decrease when controlling for size and centrality. Panel B replaces total investments with investments in hub areas. The results are largely similar in both significance and magnitude to those in Panel A: VCs that hire venture partners increase their number of investments in hub areas. In Panel C, there is no significant increase in non-hub investments, and in Panel D, which focuses on investments in low-activity areas, the results are insignificant and have a negative sign. These findings indicates that the hiring of a venture partner, similar to the hiring of a data scientist, correlates with overall firm growth. However, unlike data scientists, venture partners drive increased investments predominantly within established hub areas rather than expanding into non-hub or low-activity regions. This distinction underscores the geographically concentrated nature of traditional VC networks, highlighting the unique potential of data technologies to identify promising startups beyond major hubs, thus expanding the reach of venture capital into previously underserved locations.

5.1.2 Selection: Instrumental Variables Approach

The previous section presented evidence that growth is not the primary driver of increased investments in non hub locations. However, to address the possibility of other omitted variables, I develop an empirical strategy to estimate the causal impact of data technology adoption on VC investments. My approach isolates variation in VCs' adoption of data technologies that stems from early exposure to AI and the timing of raising a new fund, mitigating potential bias from demand shocks that could influence both technology adoption and investment strategies.

Identification Strategy I construct my instrument by isolating variation in VCs' adoption of data technologies from two sources: early exposure to AI and the timing of raising a new fund. For the early exposure measure, the intuition is that commercial interest in AI only became widespread around 2010, when technology firms first began incorporating AI into consumer products (e.g., Apple introducing Siri in 2011), followed by non-technology firms using AI to enhance business operations (e.g., Walmart deploying cameras on floor scrubbers to track real-time inventory levels in 2017). In 2012, researchers at Google introduced a deep Convolutional Neural Network (CNN) architecture that won the ImageNet challenge, sparking a surge in deep learning research and implementation (Krizhevsky, Sutskever, and Hinton (2012)). Since then, many firms — both in and outside the tech sector — have adopted these advances in their operations.

Recent research by Babina et al. (2024) highlights a significant increase in AI investments by public firms across industries, leading to growth in sales, employment, and market valuations. For early AI adopters in industries other than technology, such as VCs in the financial services industry, firms need both an understanding of the benefits of AI and the know-how to implement it. While AI became popular for commercial use after 2010, young, innovative startups were pioneering AI development in the 2000s. For instance, Predictix, founded in 2005, provided clients with big data and analytics processes to forecast business operations, and Voci, founded in 2008, pioneered speech-to-text algorithms in hardware. VCs that invested in such startups would have had a first-mover advantage in understanding AI's potential applications, ahead of other investors. I hypothesize that VCs who invested in AI-focused startups before 2010 are more likely to be early adopters of data technologies and to adjust their investment strategies in line with my previous findings.

AI Industry Exposure I exploit the cross-sectional heterogeneity in the impact of AI industries to identify the effect on data technology adoption by VCs. Crunchbase categorizes companies into 750 industries to account for heterogeneity across startup's specific market segments[11]. Following methodology used in Bonelli (2023), I assign a treatment intensity to each industry in Crunchbase proxying for the extent to which that industry would specialize in artificial intelligence. To create industry-level treatment intensities, I rely on business descriptions of firms in the Crunchbase database, including those of firms that were not VC-funded (Crunchbase covers other types of firms - including public and private that are \were not necessarily VC-backed). I start by collecting AI terms defined in the Artificial Intelligence Glossary from Tech Target, a marketing company that provides data-driven services to business-to-business technology vendors[12]. Table A4 reports the terms contained in the glossary. They include keywords such as "Artificial Intelligence", "Machine Learning" and "Natural Language Processing". I then search for these terms in the business descriptions of all companies in Crunchbase[13]. Finally, for each industry I compute the fraction of company descriptions featuring at least one AI term and I rank industries according to this metric. I only consider industries with more than 100 business descriptions to avoid assigning industry-level treatment intensities that are too dependent on a few companies. Treatment intensity

[11] For example, in the market segment *Financial Services*, Crunchbase includes Life Insurance, FinTech, Mobile Payments, and Wealth Management as some of the industries

[12] See https://www.techtarget.com/whatis/feature/Artificial-intelligence-glossary-60-terms-to-know

[13] I exclude firms classified only as "investors"

(between 0 and 1) is then defined as the overall percentile rank in the industry distribution:

$$IndustryExposure_i = \text{Rank}_I\{\frac{\text{Nb. Company Descriptions with Match in Industry } i}{\text{Nb. Company Descriptions in Industry } i}\} \quad (4)$$

where I is the set of industries in Crunchbase. Intuitively, industries in which companies mention AI terms more often are more likely to be part of the AI industry. Panel A of Table A5 shows the ten industries with the highest treatment intensities. It includes industries such as "Machine Learning", "Artificial Intelligence", "Natural Language Processing", and "Text Analytics". The least exposed industries are presented in Panel B and encompass industries such as "Timber", "Bakery", and "Laundry". This is not surprising as companies in these industries are less likely to benefit from AI.

VC Exposure The extent to which VCs are exposed to AI pre-2010 depends on the VCs sectoral specialization. A VC firm mainly investing in software and data analytics companies is more likely to invest in a firm in the AI industry. By contrast, a VC firm investing in pharmaceuticals is less likely to invest in firms conducting business in AI. My empirical strategy makes use of these variations across VC firms to identify the impact of investing in AI startups pre-2010 on VCs adoption of data technologies. An important assumptions is that VCs investing in AI startups prior to 2010 did not do so in anticipation to adopt these technologies themselves. However, this runs counter to the lack of commercial interest in AI by firms prior to 2010, especially in the non-technology sector (such as VCs in the Financial Services industry). To quantify a VC firm's exposure to the AI industry pre-2010, I create a measure called "VC Exposure" constructed by linking each VC investment in my sample to the corresponding industry exposure defined above. This creates the following exposure measure:

$$VCExposure_j = \frac{1}{N_{j,2010}} \sum_{i \in A_{j,2010}} IndustryExposure_i, \quad (5)$$

where J is the set of VCs with investments before 2010, $A_{j,2010}$ is the set of investments made by VC firm j before 2010, $N_j 2010$ is the number of investments in this set, $IndustryExposure_i$ is the treatment intensity of the industry of the startup corresponding to investment i, defined in Equation 4. VC firms with the highest exposure are those with most of their investments before 2010 in industries with high treatment intensity, creating within-industry variations across investments made by investors with different VC-level exposures.

Fund Timing Lastly, I create an indicator if a VC raised a fund in the previous two years ($NewFund_{j,[-2:0]}$). The intuition is that VCs tend to hire new employees from the increased capital they receive from raising a new fund. Thus VCs that plan on adopting more data-driven approaches are likely to wait till their next fundraising cycle. I collect data for VC fund vintages from Preqin.

First Stage I instrument VCs' data technology adoption with their exposure to AI prior to 2010 interacted with whether they raised a new fund in the previous two years. The exclusion restriction is satisfied in that commercial interest in AI for non-technology firms only became popular after 2010 and thus any investments in AI prior to 2010 were not in anticipation to adopt these technologies. To further support this assumption, the first investment made by a data-driven VC was in 2010. The following is the first-stage specification:

$$DataDriven_{j,t} = \beta VCExposure_j \times NewFund_{j,[-2:0]} + X_{j,t} + \alpha_j + \alpha_{c \times i \times s \times t} + \epsilon_{j,t}, \quad (6)$$

where $DataDriven_{j,t}$ is an indicator if the VC is data driven as of year t. $VCExposure_j$ is a VCs' exposure to AI through their investment prior to 2010 as defined in Equation 5. $NewFund_{j,[-2:0]}$ is an indicator if a VC raised a new fund in the previous two years. $X_{j,t}$ are time varying VC controls. α_j are VC-firm fixed effects and $\alpha_{c \times i \times s \times t}$ are VC state × industry × stage × year fixed effects. Standard errors are clustered at the VC firm level.

The results of the first stage are displayed in column (1) of Table 6. The coefficient on $VCExposure_j$ is positive and statistically significant and the F-statistic 14, greater than the conventional level of 10, satisfying the relevance condition for the instrument.

Second Stage Next, I implement the second stage of my instrumental specification. I estimate the following regression:

$$Y_{j,t} = \beta Data\hat{D}riven + X_{j,t} + \lambda_j + \lambda_{c \times i \times s \times t} + \xi_{j,t}, \quad (7)$$

where $Data\hat{D}riven$ is instrumented by VCs' exposure to AI prior to 2010 and $Y_{j,t}$ is the number of investments made by VC j in year t. The empirical specifications in Equation 6 and Equation 7 require observing the industry composition of VCs portfolios before 2010. This analysis therefore consists of 3,301 VC-years made by 398 VC firms. The summary statistics for this sample can be found in Panel E of Table 1. The even columns in Table 6 show the OLS results using this sample. The results are similar to that of the baseline specification in Table 3.

The results for the second stage can be found in Column (3) for investments in non-hubs and Column (5) for investments in low activity commuting zones. Since I perform Poisson Pseudo Maximum Likelihood models, I bootstrap the standard errors over 1,000 iterations. The coefficients are positive and statistically significant across both specifications, implying a causal effect of data technology adoption on lowering search frictions and VCs investing more outside of VC hubs. The magnitudes in the second stage are significantly larger than the OLS estimates. One potential reason is that only 4% of firm-years are data-driven. As a result, there may not be enough variation in the instrument, resulting in the model overfitting the data.

5.2 Other Proxies for Out-of-Network Investments

In this section, use other proxies for startups considered to be outside VC networks to test where data technology adoption lowers search frictions for finding investments. I conduct these analyses at the investment level, to focus on VC-startup pairs, using the following regression:

$$Y_{j,k,t} = \beta DataDriven_{j,t} + X_{j,k,t} + \alpha_j + \alpha_{c\times i\times s\times t} + \epsilon_{j,k,t} \tag{8}$$

For my first test I investigate whether VCs invest further away after adopting data technologies. Specifically, the dependent variable is an indicator if the investment is located in the top tercile of distance over my sample period. The main explanatory variable, $DataDriven$, is a dummy variable equal to 1 if VC j is classified as data-driven as of the investment date and 0 otherwise. $X_{j,k,t}$ includes the same set of time varying controls for VC j and also includes a rich set of time varying startup k, and VC j startup k pair controls. Specifically, I control for startup age ($Log(StartupAge)$), where the founder is serial entrepreneur ($Serial$), whether the founder was previously a VC ($VCPrior$), and whether a partner at VC j and a founder at startup k share the same alma mater ($Alumni$). I estimate Equation 8 and Column (1) of Table 7 displays the results without the VC firm fixed effect. The coefficient is highly insignificant, indicating that data-driven VCs are no more likely to invest in distantly located startups than traditional VCs. In column (2), I include the VC-firm fixed effect. While still insignificant, the coefficient is positive and can be interpreted as, after VCs adopt data technologies they are 6% more likely to invest in distantly located startups.

In columns (3) and (4), I replace the outcome variable with an indicator if the VC invests in a different industry than their specialization. A large literature shows that VCs tend to specialize in investing in various industries (e.g. Hochberg, Mazzeo, and McDevitt (2015)) and these industries can form established networks within the VC industry (Hochberg, Ljungqvist, and Lu (2010)). I therefore classify a VC as specializing in a particular industry if more than 40% of their investments in the previous 5 years are in startups from the same industry. Crunchbase uses a granular industry specification system with over 750 industry classifications. Using a supervised machine learning approach, I classify these into 7 industry groups (Figure A1): Software and IT, Health Care and Biotechnology, Hardware and Electronics, Financial Services, Business Services, Consumers, Industrial and Energy. The coefficient on $DataDriven$ in columns (3) and (4) is positive and statistically significant, and can be interpreted as, after VCs adopt data technologies they are 7.7% more likely to invest in a startup in a different industry than their specialization, a 40% increase from the unconditional mean.

Lastly, in columns (5) and (6), I investigate whether VCs rely less on other investors to find startups and invest with. I therefore replace the outcome variable with an indicator equal to 1 if a VC invests with another VC located in the same state as the startup, conditional on investing out of their headquartered state. The intuition is that VCs who invest outside of their home state are less likely to know of potential investment opportunities in other VC markets unless they know another VC located close to the startup. However, if VCs use data technologies to find investments, they can now find the startup without local help. The coefficients on $DataDriven$ are negative and statistically significant in Column (6) and can be interpreted as, after VCs adopt data technologies, and conditional on investing outside of their headquartered state, they are 2% less likely to syndicate with a local VC, a

6 Data-Driven Performance

In the previous section, I find that VCs increase the overall quantity of investments after adopting data technologies, specifically in non-hub and low-activity areas. In this next section, I examine the quality of these data-driven investments. Prior literature shows that data technology adoption provides an advantage in finding startups that are less likely to fail (i.e. more likely to receive follow on funding) but provides no advantage in finding firms that are likely to achieve a major exit, either through an IPO or an acquisition (Bonelli (2023)). However, little is known how data-driven investments in non hub and low activity areas compare to (1) their data-driven hub counterparts or (2) traditional investments in non hub and low activity areas. Ex ante, the performance of data-driven non-hub investments is unclear. On one hand, VCs investing in non-hubs may encounter less competition, potentially enabling them to invest in higher quality startups. This aligns with prior literature that VCs have a higher hurdle rate for their non hub investments, and as a result, those investments tend to outperform their local counterparts (Chen et al. (2010)). Alternatively, if data technologies reduce search frictions for finding investments, they may identify more lower quality firms in non hub areas. VCs actively monitor their portfolio companies, which leads to increased performance (Bernstein, Giroud, and Townsend (2016)). However, since non hub investments are typically further away, VCs may be unable to effectively manage their lower quality investments. Thus the performance of data-driven non-hub and low-activity investments is an empirical question.

I therefore examine the performance of data-driven investments compared to traditional VCs. I split performance outcomes into three variables: whether the startup received follow-on funding, if the startup achieved an exit through a high-quality acquisition, or if the startup exited through an IPO or achieved Unicorn status. I examine this at the investment level and replace the outcome variable in Equation 8 with an indicator if the startup receives follow on funding, exits through an acquisition or exits through an IPO or achieves unicorn status.

First, I investigate if data-driven investments are more likely to receive follow-on funding. I restrict my sample to all investments made between 2010 and 2020, as investments take up to two years to receive their next investment. Prior literature finds that overall, data-driven investments are more likely to receive follow-on funding. I find the same result in column (1) of Table 8. The coefficient on $DataDriven$ can be interpreted as, after VCs adopt data technologies, the likelihood their investments receive follow on funding increases by 2.6 percentage points or 3.3% compared to the unconditional mean. In column (2) I interact $DataDriven$ with an indicator equal to 1 if the startup is located in a non-hub. The coefficient on the interaction term is negative, but statistically insignificant. The coefficient on $DataDriven$ is still positive and statistically significant and the coefficient between the interaction and $DataDriven$ is statistically different at the 5% level, indicating that data-driven investments in non-hub locations are less likely to receive a follow-on investment than their data-driven counterparts in hub locations. I also compare the coefficient on the interaction term to the coefficient on $NonHub$, which represents traditional investments in non-hub areas. The difference is not statistically different, indicating that data-driven investments in non-hub areas do not perform any differently to traditional investments in non-hubs. In column (3), I conduct the same analysis but replace non-hub with an indicator if the startup is located in a low activity area. The coefficient is negative and statistically significant, indicating that data-driven low-activity investments are less likely to receive follow-on funding. They are also significantly less likely to receive follow-on funding than their non low-activity data-driven counterparts ($DataDriven$) and traditional investors in low-activity areas ($LowActivity$). In columns (4) through (6), I repeat the same analysis but restrict investments to those where the investor is the lead VC in that funding round. The results are similar, with the coefficients slightly larger in magnitude, indicating that overall, VCs lead investments are more likely to receive follow-on funding, except in non-hub and low-activity areas.

The above analysis suggests that data-driven VCs tend to invest in lower quality startups in non-hub and low-activity areas than their hub investments. However, VCs make the majority of their returns through successful exit events, primarily IPOs and acquisitions. These exit routes allow VCs to capitalize on the growth and value of their portfolio companies, translating their equity stakes into liquidity. Among these exits, IPOs are generally regarded as the most lucrative outcome, generating a 5x to 10x or even 20x return, versus an acquisition, which tends to yield a 1x to 5x return. However, over the last decade, firms are staying private longer (Ewens, Nanda, and Rhodes-Kropf (2018)) and are often much older and larger when they eventually do go public (Gao, Ritter, and Zhu (2013)). Successful startups are likely to achieve unicorn status (a valuation above $1 billion) before going public. In addition, it is rare for startups to achieve such a return. Only 23% of startups in my sample exit through an acquisition, 7% through an IPO, and only 10% achieve unicorn status. Thus VCs tend to make the majority of their returns in the right tail of their return distribution. To test this, I start by investigating whether VCs are more likely to invest in startups that eventually IPO or receive unicorn status after they adopt data technologies. I restrict my sample to first-time investments made by VCs between 2010 to 2018 as startups typically take 4-5 years to exit. The results are displayed in Table 9.

The statistically insignificant coefficient on $DataDriven$ in column (1) supports prior literature's findings that data technologies have no effect on identifying firms that are likely achieve a major exit through an IPO or achieve unicorn status. However, in column (2), the coefficient on the interaction term of $DataDriven$ and $NonHub$ is positive and statistically significant, indicating that the likelihood of data-driven investments exiting through an IPO or reaching unicorn status increases by 2.6 percentage points, an increase of 26%. The coefficients on $DataDriven$ and $DataDriven \times NonHub$ are also statistically different from one another, indicating that data-driven investments in non hubs are more likely to exit through IPO or achieve unicorn status than their data-driven hub counterparts. Similarly, the coefficient on $NonHub$ and $DataDriven \times NonHub$ is statistically different from one another, indicating that data-driven investments in non-hubs are more likely to outperform than traditional investments in non-hubs. I find similar results in column (3) when I replace $NonHub$ with $LowActivity$. The coefficient magnitudes are even larger, and can be interpreted as the likelihood of data-driven investments in low-activity exiting through an IPO or achieving unicorn status increases by 4.5 percentage points or by 56% compared to the unconditional mean. The coefficient magnitudes continue to increase in columns (5) and (6) where I further restrict the analysis to VCs' lead investments.

To test whether data technologies provide an advantage for finding investments that are likely to exit through an acquisition, I repeat the same analysis, but replace the dependent variable with an indicator if the startup exits through an acquisition. The results are displayed in Table 10. In all specifications, the likelihood that a startup exits through an acquisition is no different after VCs adopt data technologies (columns (1) and (4)), regardless if the startup is located in a non hub ((2) and (5)) or low activity ((3) and (6)) area.

Overall, the results suggest that data-driven technologies are able to identify high-quality startups that are more likely to achieve unicorn status or exit through an IPO, but also identify startups in these areas that are more likely to fail (and, also, has no impact on acquisition outcomes in non-hub areas). These results can be interpreted in two ways. Either, data technologies lower search frictions for finding startups in non hub areas, both of higher quality (as shown through the IPO and unicorn analysis) and lower quality (as shown through the follow-on analysis) than their hub investments. Alternatively, VCs are more likely to abandon lower quality startups in non hub areas in earlier rounds than their hub counterparts. VCs are better able to monitor their local investments or investments where they are more likely to have a secondary office (i.e. another hub location). Thus they may continue to invest in startups in hub areas even if they are of lower quality, as the monitoring may offset their concerns. In either interpretation, VCs make the majority of their returns through an IPO or acquisition exit, not through their startup receiving another funding round. Thus data technologies not only identify startups in non-hub locations, but they are able to identify better quality startups than traditional VC investments in non-hub locations, as well as better quality investments than data-driven investments in hub locations.

7 Which Non-Hub Areas Are Likely to Attract Data-Driven Investments?

To summarize the results so far, I show that after VCs adopt data technologies, they are more likely to invest in non-hub and low activity commuting zones, and that these investments have a higher likelihood of exiting through and acquisition or IPO. In this next section, I investigate which commuting zones are most likely to attract data-driven investments. One advantage of algorithmic techniques is that are able to identify emerging trends and markets. I therefore posit that data-driven VCs are more likely to identify startups in areas with high levels of entrepreneurial activity. I first describe the data and research design and then present my findings.

7.1 Data and Research Design

I begin by constructing a commuting-zone-year panel from 2010 to 2017[14]. To identify the entrepreneurial activity of a commuting zone, I use recently available data from the "Startup Cartography Project" or SCP (Fazio et al. (2019)). The SCP offers a set of entrepreneurial ecosystem statistics for the United States at the zip code, county, MSA, and state level from 1988 to 2016. The SCP combines state-level business registration records with a predictive analytics approach to estimate the probability of "extreme" growth (IPO or high-value acquisition) at or near the time of founding for all newly registered firms in a given year. The SCP then leverages estimates of entrepreneurial quality to develop four entrepreneurial ecosystem statistics, including the rate of start-up formation, average entrepreneurial quality, the quality-adjusted quantity of entrepreneurship, and entrepreneurial ecosystem performance over time. For my analysis, I use their *Regional Entrepreneurship Cohort Potential Index* or $RECPI$. This measure interacts the number of new business registrants within a given population with the average growth potential or quality of those startups. In short, the $RECPI$ is a quality-adjusted index for entrepreneurial activity in a given area. Summary statics can be found in Panel D of Table 1. I construct the following commuting-zone-year panel:

$$\#Investments_{c,t} = \beta Log(RECPI)_{c,t-1} + X_{c,t-1} + \alpha_c + \alpha_t + \epsilon_{c,t}, \tag{9}$$

The main dependent variable is $\#Investments_{c,t}$, the number of VC investments made in commuting zone c in year t. $Log(RECPI)_{c,t-1}$ is the natural log of $RECPI$ of commuting

[14] The data from the Startup Cartography Project ends in 2016.

zone c in the previous year. $X_{c,t-1}$ includes time varying commuting zone controls such as GDP, Income, and the percentage of the adult population that went to college. Since the number of investments is a count variable with certain specifications left-censored at zero and skewed, I estimate a Poisson model. Standard errors are clustered at the commuting zone level. My main prediction is that data technologies are skilled in identifying emerging markets and trends, and thus commuting zones with higher levels of $RECPI$ will attract more data-driven investments.

7.2 Results

The results are displayed in Table 11. Panel A shows the number of data-driven investments. The odd numbered columns do not include commuting zone fixed effects and thus represent the impact of high-levels of entrepreneurial activity on attracting data-driven investments. Column (1) includes all commuting zones. The coefficient on $Log(RECPI)$ is positive and statistically significant and can be interpreted as a one standard deviation increase in $Log(RECPI)$ results in a 632.4% increase in the number of data-driven investments. In column (3) I exclude all major VC-hub years and repeat the analysis. The coefficient, while slightly smaller in magnitude, is still positive and statistically significant, indicating that areas outside major hubs but with high levels of entrepreneurial activity attract more data-driven investment. I further restrict my sample in column (5) by including on commuting zones with low VC activity (i.e. 25 or fewer VC investments in the prior 5 years). While still smaller in magnitude still, the coefficient is positive and statistically significant, indicating that areas with low venture activity but high entrepreneurial activity attract more data-driven investments.

The even numbered columns in Table 11 include commuting-zone fixed effects. This removes all time invariant characteristics and thus regressing the number of investments on the lag of $Log(RECPI)$ captures the change in entrepreneurial activity. In column (2) of Panel A, the coefficient is negative and statistically significant, indicating that changes in $Log(RECPI)$ lead to a decrease in data-driven investment. However, I find the result reverses when I exclude major VC-hub years in column (4). The coefficient magnitude is even higher when further restricting the sample to areas with low VC activity. The negative result in column (2) may be due to VC hubs having high levels of $Log(RECPI)$ and VC capital investment, and since the data is highly skewed, including the commuting-zone fixed effect concentrates the results in these areas. Overall, these results can be interpreted as data-driven VCs are more likely to invest in areas with growing entrepreneurial activity, which can be attributed to data technologies identifying emerging market trends.

I repeat the analysis in Panel B of Table 11, however the dependent variable is the number of investments made by traditional VCs in a commuting-zone year. For columns (1), (3), and (5), the results are similar in significance and magnitude to that of data-driven investments. However, when I include the commuting zone fixed effects in columns (2), (4), and (6), the results become insignificant and the magnitudes decrease drastically. This suggests that traditional methods of identifying startups may overlook emerging markets. These findings further support the notion that data technologies are more effective in identifying emerging markets and trends, offering a valuable tool for VCs looking to expand into these areas.

7.3 Implications for Areas with Low History of VC Activity

In the previous section, I demonstrated that non hub and low activity areas with growing entrepreneurial activity attract more data-driven investment. Next, I explore whether investment in these areas, specifically low activity areas, experience an increase in subsequent VC activity. I begin by constructing a panel of all low activity commuting zones in the US during my sample period (i.e. 25 or fewer VC investments over the previous five years Hochberg, Ljungqvist, and Lu (2010)). I then identify startups in these commuting zones that receive funding for the first time by data-driven VCs and classify these commuting zones as treated, a total of 56 commuting zones. I classify all other commuting zones as my control group. I then construct a stacked difference-in-difference model, comparing various measures of VC activity before and after an investment made in the commuting zone by a data-driven VC. Specifically, I construct the following difference-in-difference regression:

$$Y_{d,c,t} = \beta\{Treated_{d,c} \times Post_{d,t}\} + \alpha_{d,c} + \alpha_{d,t} + \epsilon_{d,c,t}, \tag{10}$$

where $Y_{d,c,t}$ are various outcomes of venture activity for commuting zone c in cohort d and year t. $Treated_{d,c}$ is an indicator equal to one if a startup in commuting zone c received an investment by a data-driven VC. $Post_{d,t}$ is an indicator that equals one post data-driven entry and zero otherwise. The baseline specification controls for cohort × county ($\alpha_{d,c}$) to absorb any time-invariant characteristics at the commuting zone level and and cohort × year ($\alpha_{d,t}$) fixed effects to absorb time trends. In addition, I add pre-data-driven entry commuting zone characteristics including income, gdp, and percentage of the population that has a college degree, interacted with $Post_{t,c}$ to account for the possibility that commuting zones with certain characteristics experience a change in outcomes post data-driven entry. In the tightest specification, I also include $Log(RECPI)$ interacted with $Post$ to control for the level of entrepreneurial activity prior to investment. All outcomes are left-censored at zero and skewed and therefore I estimate a Poisson model. The variable of interest , β, captures the change in an outcome variable for commuting zones with a data-driven investment ($Treated_{d,c}$) to those without.

I begin by looking at the impact of an investment by a data-driven VC in a commuting zone with low VC activity on entrepreneurial activity in subsequent years. Specifically, I look at the number funding rounds, the number of startups that receive their first ever VC financing, the number of unique investors, the number of first time investors and the number of patents filed by startups backed by VCs. I following methodology introduced by Ewens and Marx (2024) to classify these patents as being filed by VC-backed startups. I run the specification outlined in Equation 10. The results are displayed in Table 12. In columns (1) and (2), I find that the number of funding rounds increases by 8-12% in commuting zones that experience an investment by a data-driven VC compared to commuting zones that do not. Similarly in columns (3) and (4), I find that the number of startups that receive their first ever VC-financing also increase by 8-12%. In columns (5) and (6) I find that the number of unique investors in low activity commuting zones that receive data-driven investment experience an increase of 19-26% and the in columns (7) and (8) I find that the number of first-time investors increases by 11-18%. Lastly, in columns (9) and (10), the number of patents produced by VC-backed startups increases by approximately 28% to 35% in commuting zones that receive investment from a data-driven VC. Overall, the increase in VC activity in low activity commuting zones after entry by a data-driven investor indicates that data technologies can have a positive impact on the financing of innovation in areas outside major clusters in the US.

8 Conclusion

The adoption of data technologies by VCs firms has the potential to significantly transform their investment strategies and the broader landscape of innovation. This paper demonstrates that data technologies enable VCs to broaden their investment opportunity sets, allowing them to identify and invest in startups beyond their traditional networks and geographic constraints. By leveraging detailed employee data from Crunchbase and LinkedIn, I track the adoption of data technologies and show that VCs become more likely to invest outside major hubs, specifically those with high and growing levels of entrepreneurial activity. In addition, data-driven investments in non hub locations are more likely to IPO or achieve unicorn status than their hub counterparts. I also find that areas with little prior VC activity that receive a data-driven investment experience an increase in subsequent VC activity. These findings suggest that data-driven approaches can mitigate information frictions and enhance the efficiency of deal sourcing. As data technologies continue to evolve, their role in democratizing access to venture capital and consequently impacting entrepreneurial growth in underrepresented areas will likely become increasingly important. Future research should continue to explore the long-term effects of this technological shift on financial markets.

References

Abis, Simona. 2020. "Man vs. machine: Quantitative and discretionary equity management." *Machine: Quantitative and Discretionary Equity Management (October 23, 2020).*

Abis, Simona, and Laura Veldkamp. 2024. "The changing economics of knowledge production." *The Review of Financial Studies* 37 (1): 89–118.

Alekseeva, Liudmila, José Azar, Mireia Gine, Sampsa Samila, and Bledi Taska. 2021. "The demand for AI skills in the labor market." *Labour economics* 71:102002.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. 2024. "Artificial intelligence, firm growth, and product innovation." *Journal of Financial Economics* 151:103745.

Bai, Jennie, Thomas Philippon, and Alexi Savov. 2016. "Have financial markets become more informative?" *Journal of Financial Economics* 122 (3): 625–654.

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144 (2): 370–395.

Bernstein, Shai, Xavier Giroud, and Richard R Townsend. 2016. "The impact of venture capital monitoring." *The Journal of Finance* 71 (4): 1591–1622.

Birru, Justin, Sinan Gokkaya, and Xi Liu. 2018. *Capital market anomalies and quantitative research.* Technical report.

Blattner, Laura, and Scott Nelson. 2021. "How costly is noise? Data and disparities in consumer credit." *arXiv preprint arXiv:2105.07554.*

Bonelli, Maxime. 2023. "Data-driven Investors." In *Data-driven Investors: Bonelli, Maxime.* [Sl]: SSRN.

Cao, Sean, Wei Jiang, Junbo Wang, and Baozhong Yang. 2024. "From man vs. machine to man+ machine: The art and AI of stock analyses." *Journal of Financial Economics* 160:103910.

Chen, Henry, Paul Gompers, Anna Kovner, and Josh Lerner. 2010. "Buy local? The geography of venture capital." *Journal of Urban Economics* 67 (1): 90–102.

Chi, Feng, Byoung-Hyoun Hwang, and Yaping Zheng. 2023. "The Use and Usefulness of Big Data in Finance: Evidence from Financial Analysts." *Nanyang Business School Research Paper,* nos. 22-01.

Cohen, Lauren, Andrea Frazzini, and Christopher Malloy. 2008. "The small world of investing: Board connections and mutual fund returns." *Journal of Political Economy* 116 (5): 951–979.

Coleman, B, KJ Merkley, and J Pacelli. 2021. "Do robot analysts outperform traditional research analysts." *The Accounting Review, forthcoming.*

Coval, Joshua D, and Tobias J Moskowitz. 2001. "The geography of investment: Informed trading and asset prices." *Journal of political Economy* 109 (4): 811–841.

Cumming, Douglas, and Na Dai. 2010. "Local bias in venture capital investments." *Journal of empirical finance* 17 (3): 362–380.

DâAcunto, Francesco, Nagpurnanand Prabhala, and Alberto G Rossi. 2019. "The promises and pitfalls of robo-advising." *The Review of Financial Studies* 32 (5): 1983–2020.

Davenport, Diag. 2022. "Predictably bad investments: Evidence from venture capitalists." *Available at SSRN 4135861.*

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard. 2021. "Does alternative data improve financial forecasting? the horizon effect."

Dessein, Wouter. 2005. "Information and control in ventures and alliances." *The Journal of Finance* 60 (5): 2513–2549.

Di Maggio, Marco, Dimuthu Ratnadiwakara, and Don Carmichael. 2022. *Invisible primes: Fintech lending with alternative data.* Technical report. National Bureau of Economic Research.

Dugast, Jérôme, and Thierry Foucault. 2018. “Data abundance and asset price informativeness.” *Journal of Financial economics* 130 (2): 367–391.

Ewens, Michael, Ramana Nanda, and Matthew Rhodes-Kropf. 2018. “Cost of experimentation and the evolution of venture capital.” *Journal of Financial Economics* 128 (3): 422–442.

Farboodi, Maryam, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran. 2022. “Where has all the data gone?” *The Review of Financial Studies* 35 (7): 3101–3138.

Farboodi, Maryam, and Laura Veldkamp. 2020. “Long-run growth of financial data technology.” *American Economic Review* 110 (8): 2485–2523.

Fazio, Catherine, Scott Stern, Jorge Guzman, Yupeng Liu, and RJ Andrews. 2019. “The startup cartography project: measuring and mapping entrepreneurial ecosystems.” *Research Policy.*

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. “Predictably unequal? The effects of machine learning on credit markets.” *The Journal of Finance* 77 (1): 5–47.

Gao, Meng, and Jiekun Huang. 2020. “Informing the market: The effect of modern information technologies on information production.” *The Review of Financial Studies* 33 (4): 1367–1411.

Gao, Xiaohui, Jay R Ritter, and Zhongyan Zhu. 2013. “Where have all the IPOs gone?” *Journal of Financial and Quantitative Analysis* 48 (6): 1663–1692.

Garfinkel, Jon A, Erik J Mayer, Ilya A Strebulaev, and Emmanuel Yimfor. 2021. “Alumni networks in venture capital financing.” *SMU Cox School of Business Research Paper,* nos. 21-17.

Glaeser, Edward L, William R Kerr, and Giacomo AM Ponzetto. 2010. "Clusters of entrepreneurship." *Journal of urban economics* 67 (1): 150–168.

Goldfarb, A, B Taska, and F Teodoridis. 2021. "Could machine learning be a general purpose technology." *A comparison of emerging technologies using data from online job postings.*

Gompers, Paul A, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev. 2020. "How do venture capitalists make decisions?" *Journal of Financial Economics* 135 (1): 169–190.

Gompers, Paul A, Vladimir Mukharlyamov, Emily Weisburst, and Yuhai Xuan. 2022. "Gender gaps in venture capital performance." *Journal of Financial and Quantitative Analysis* 57 (2): 485–513.

Gormley, Todd A, and David A Matsa. 2011. "Growing out of trouble? Corporate responses to liability risk." *The Review of Financial Studies* 24 (8): 2781–2821.

Gornall, Will, and Ilya A Strebulaev. 2021. "The economic impact of venture capital: Evidence from public companies." *Available at SSRN 2681841.*

Grennan, Jillian, and Roni Michaely. 2020. "Artificial intelligence and high-skilled work: Evidence from analysts." *Swiss Finance Institute Research Paper,* nos. 20-84.

Grossman, Sanford J, and Joseph E Stiglitz. 1980. "On the impossibility of informationally efficient markets." *The American economic review* 70 (3): 393–408.

Hochberg, Yael V, Alexander Ljungqvist, and Yang Lu. 2007. "Whom you know matters: Venture capital networks and investment performance." *The Journal of Finance* 62 (1): 251–301.

———. 2010. "Networking as a barrier to entry and the competitive supply of venture capital." *The Journal of Finance* 65 (3): 829–859.

Hochberg, Yael V, Michael J Mazzeo, and Ryan C McDevitt. 2015. "Specialization and competition in the venture capital industry." *Review of Industrial Organization* 46:323–347.

Hong, Harrison, and Jiangmin Xu. 2019. "Inferring latent social networks from stock holdings." *Journal of Financial Economics* 131 (2): 323–344.

Howell, Sabrina T, and Ramana Nanda. 2019. "Networking frictions in venture capital, and the gender gap in entrepreneurship." *Journal of Financial and Quantitative Analysis,* 1–56.

Hsu, David H. 2004. "What do entrepreneurs pay for venture capital affiliation?" *The journal of finance* 59 (4): 1805–1844.

Huang, Can. 2022. "Networks in venture capital markets." *Available at SSRN 4501902.*

Huberman, Gur. 2001. "Familiarity breeds investment." *The Review of Financial Studies* 14 (3): 659–680.

Kaplan, Steven N, Berk A Sensoy, and Per Strömberg. 2009. "Should investors bet on the jockey or the horse? Evidence from the evolution of firms from early business plans to public companies." *The Journal of Finance* 64 (1): 75–115.

Kaplan, Steven N, and Per Strömberg. 2000. "How do venture capitalists choose investments." *Workng Paper, University of Chicago* 121:55–93.

Koenig, Lukas. 2022. "Cut From the Same Cloth: The Role of University Affiliations in Venture Capital Investments." *Available at SSRN 4248420.*

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25.

Kuchler, Theresa, Yan Li, Lin Peng, Johannes Stroebel, and Dexin Zhou. 2022. "Social proximity to capital: Implications for investors and firms." *The Review of Financial Studies* 35 (6): 2743–2789.

Lerner, Josh, and Ramana Nanda. 2020. “Venture capitalâs role in financing innovation: What we know and how much we still need to learn.” *Journal of Economic Perspectives* 34 (3): 237–261.

Lerner, Joshua. 1994. “Venture capitalists and the decision to go public.” *Journal of financial Economics* 35 (3): 293–316.

———. 2010. *Geography, Venture Capital and Public Policy.* Rappaport Institute/Taubman Center.

———. 2022. “The syndication of venture capital investments.” In *Venture Capital,* 207–218. Routledge.

Li, Wenfei, Donghui Li, and Shijie Yang. 2022. “The impact of internet penetration on venture capital investments: Evidence from a quasi-natural experiment.” *Journal of Corporate Finance* 76:102281.

Lyonnet, Victor, and Léa H Stern. 2022. “Venture capital (mis) allocation in the age of AI.” *Fisher College of Business Working Paper,* nos. 2022-03, 002.

Massa, Massimo, and Andrei Simonov. 2006. “Hedging, familiarity and portfolio choice.” *The Review of Financial Studies* 19 (2): 633–685.

Pool, Veronika K, Noah Stoffman, and Scott E Yonker. 2012. “No place like home: Familiarity in mutual fund manager portfolio choice.” *The Review of Financial Studies* 25 (8): 2563–2599.

———. 2015. “The people in your neighborhood: Social interactions and mutual fund portfolios.” *The Journal of Finance* 70 (6): 2679–2732.

Retterath, Andre. 2020. “Human versus computer: benchmarking venture capitalists and machine learning algorithms for investment screening.” *Available at SSRN 3706119.*

Rossi, Alberto G, and Stephen P Utkus. 2020. “Who benefits from robo-advising? Evidence from machine learning.” *Evidence from Machine Learning (March 10, 2020).*

Seasholes, Mark S, and Ning Zhu. 2010. "Individual investors and local bias." *The Journal of Finance* 65 (5): 1987–2010.

Sørensen, Morten. 2007. "How smart is smart money? A two-sided matching model of venture capital." *The Journal of Finance* 62 (6): 2725–2762.

Sorenson, Olav, and Toby E Stuart. 2001. "Syndication networks and the spatial distribution of venture capital investments." *American journal of sociology* 106 (6): 1546–1588.

Van Nieuwerburgh, Stijn, and Laura Veldkamp. 2009. "Information immobility and the home bias puzzle." *The Journal of Finance* 64 (3): 1187–1215.

———. 2010. "Information acquisition and under-diversification." *The Review of Economic Studies* 77 (2): 779–805.

Verrecchia, Robert E. 1982. "Information acquisition in a noisy rational expectations economy." *Econometrica: Journal of the Econometric Society,* 1415–1430.

Zhu, Christina. 2019. "Big data as a governance mechanism." *The Review of Financial Studies* 32 (5): 2021–2061.

Figure 1: **Data-Driven Investments Over Time**

The figure reports for each year between 2000 and 2022 the number and percentage of investments made by VC firms classified as data-driven.

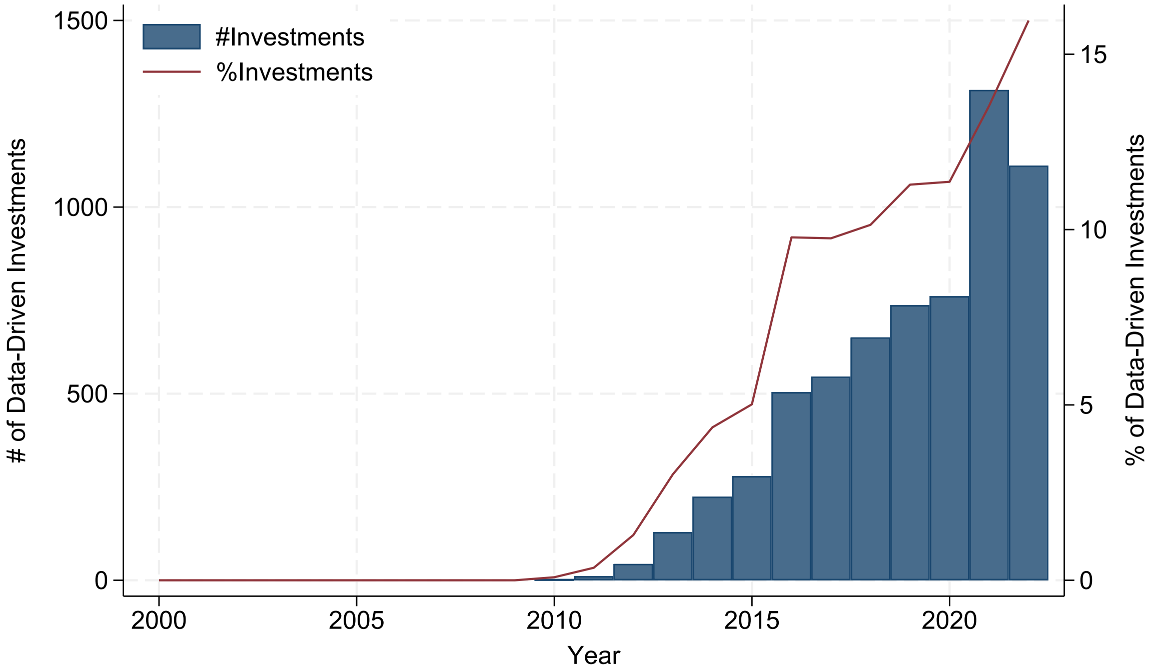
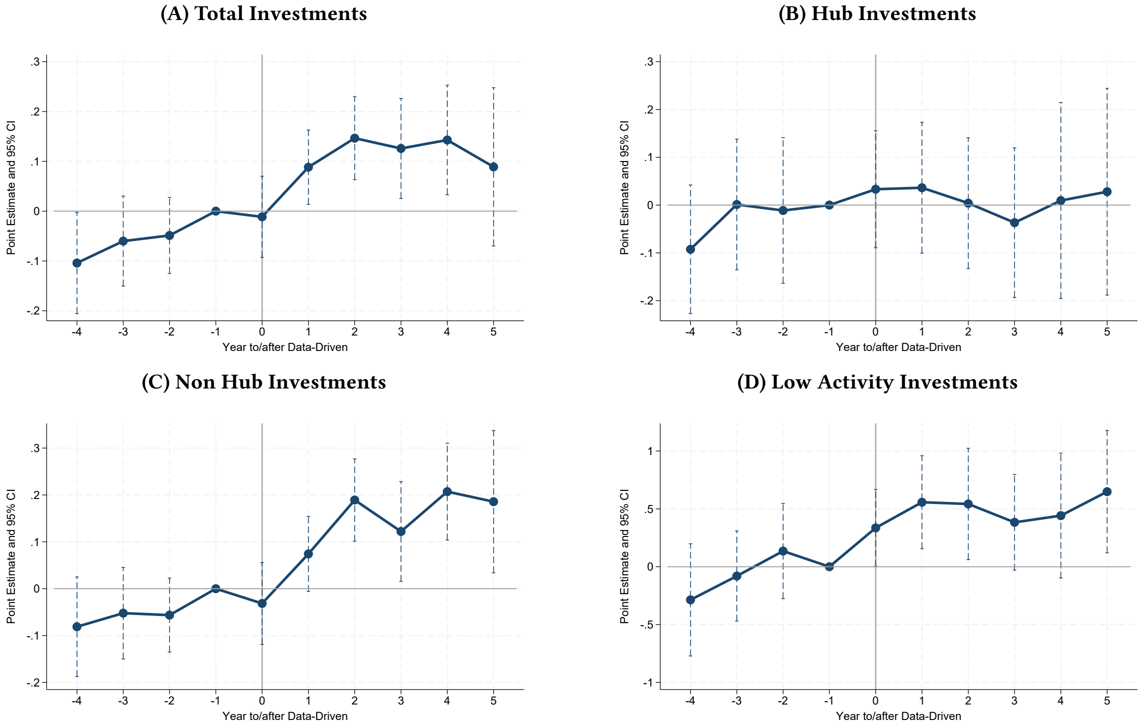


Figure 2: **Data Technologies And Number of Investments**

The figures plot the estimated coefficients from Equation 2 at the VC-Year level, of each year relative to VCs' adoption of data technologies. In Panel A, the dependent variable is the number of investments made by a VC firm in a given year. In Panel B, the dependent variable is the number of investments made in hub commuting zones by a VC firm in a given year. In Panel C, the dependent variable is the number of investments made in non hub commuting zones by a VC firm in a given year. In Panel D, the dependent variable is the number of investments made in low activity commuting zones by a VC firm in a given year. The year prior to VCs adopting data technologies is the excluded category, reported as zero in the figures. The horizontal bars represent the 90% confidence interval for the coefficient estimates with standard errors clustered at the VC firm level. Regressions include VC firm fixed effects, VCs headquarter state-main stage-main industry-year fixed effects. Regressions are conducted as Poisson Pseudo Maximum Likelihood (PPML) regressions.



(A) Total Investments

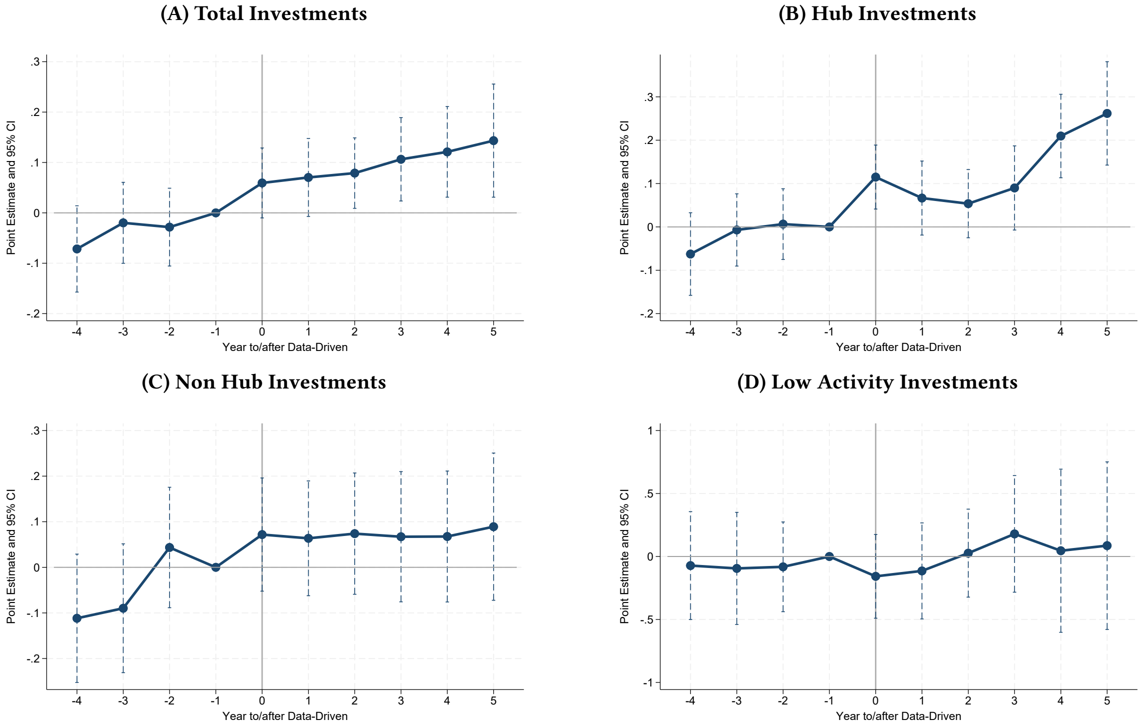
(B) Hub Investments

(C) Non Hub Investments

(D) Low Activity Investments

Figure 3: **Placebo And Number of Investments**

The figures plot the estimated coefficients from Equation 2 at the VC-Year level, of each year relative to VCs' hiring of a venture partner. In Panel A, the dependent variable is the number of investments made by a VC firm in a given year. In Panel B, the dependent variable is the number of investments made in hub commuting zones by a VC firm in a given year. In Panel C, the dependent variable is the number of investments made in non hub commuting zones by a VC firm in a given year. In Panel D, the dependent variable is the number of investments made in low activity commuting zones by a VC firm in a given year. The year prior to VCs adopting data technologies is the excluded category, reported as zero in the figures. The horizontal bars represent the 90% confidence interval for the coefficient estimates with standard errors clustered at the VC firm level. Regressions include VC firm fixed effects, VCs headquarter state-main stage-main industry-year fixed effects. Regressions are conducted as Poisson Pseudo Maximum Likelihood (PPML) regressions.



(A) Total Investments

(B) Hub Investments

(C) Non Hub Investments

(D) Low Activity Investments

Table 1: **Summary Statistics**

	Mean	St. Dev.	P1	P25	Median	P75	P99	N
Panel A: VC-Year Level								
Data Driven	0.04	0.19	0.00	0.00	0.00	0.00	0.00	8513
# Investments All	8.86	12.80	0.00	2.00	5.00	10.00	21.00	8513
# Investments First	5.45	7.69	0.00	1.00	3.00	7.00	13.00	8513
# Investments Hub All	5.95	9.96	0.00	1.00	2.00	7.00	15.00	8513
# Investments Hub First	3.58	5.81	0.00	0.00	2.00	4.00	9.00	8513
# Investments Nonhub All	2.91	4.16	0.00	0.00	2.00	4.00	7.00	8513
# Investments Nonhub First	1.87	2.82	0.00	0.00	1.00	2.00	5.00	8513
# Investments Low Comzone All	0.24	0.78	0.00	0.00	0.00	0.00	1.00	8513
# Investments Low Comzone First	0.17	0.55	0.00	0.00	0.00	0.00	1.00	8513
Investor Age	11.95	11.74	0.00	4.00	9.00	16.00	27.00	7903
Panel B: VC-Year Level Stacked								
Data Driven	0.01	0.07	0.00	0.00	0.00	0.00	0.00	34293
Venture Partner	0.18	0.23	0.00	0.00	0.00	1	3	34293
# Investments All	8.25	11.38	0.00	2.00	5.00	10.00	19.00	34293
# Investments Hub All	5.45	8.87	0.00	1.00	2.00	6.00	14.00	34293
# Investments Nonhub All	2.81	3.91	0.00	0.00	2.00	4.00	7.00	34293
# Investments Low Comzone All	0.24	0.79	0.00	0.00	0.00	0.00	1.00	34293
Investor Age t-1	0.99	1.13	0.00	0.00	0.69	1.95	2.64	34293
Panel C: Investment Level								
Data-Driven	0.06	0.25	0.00	0.00	0.00	0.00	0.00	62020
Startup Age	3.05	3.93	0.00	1.00	2.00	4.00	7.00	62020
Investor Age	14.04	15.04	0.00	4.00	9.00	20.00	33.00	62020
Distance (miles)	857	1039	0	14	283	1865	2570	49411
Local Syndicate	0.72	0.45	0.00	0.00	1.00	1.00	1.00	62020
Investment Outside Industry Specialization	0.19	0.39	0.00	0.00	0.00	0.00	1.00	49411
Follow On	0.61	0.49	0.00	0.00	1.00	1.00	1.00	62020
Acquisition	0.23	0.42	0.00	0.00	0.00	0.00	1.00	62020
IPO	0.07	0.33	0.00	0.00	0.00	0.00	1.00	62020
Unicorn	0.10	0.30	0.00	0.00	0.00	0.00	0.00	62020
Panel D: VC-Year Level - IV Sample								
Data Driven	0.05	0.22	0.00	0.00	0.00	0.00	0.00	3301
# Investments All	18.73	17.02	5.00	7.00	12.00	25.00	43.00	3301
# Investments First	9.87	9.10	0.00	4.00	6.00	13.00	23.00	3301
# Investments Hub All	13.65	15.08	0.00	4.00	8.00	19.00	34.00	3301
# Investments Hub First	7.04	8.10	0.00	2.00	4.00	10.00	18.00	3301
# Investments Nonhub All	5.42	5.28	0.00	2.00	4.00	7.00	11.00	3301
# Investments Nonhub First	3.09	3.40	0.00	1.00	2.00	4.00	7.00	3301
# Investments Low Comzone All	0.37	0.70	0.00	0.00	0.00	1.00	1.00	3301
# Investments Low Comzone First	0.26	0.71	0.00	0.00	0.00	0.00	1.00	3301
Investor Age	21.54	1.68	7.03	15.03	20.9	30.88	42.1	3301
Panel E: Commuting Zone - Year Level								
# Investments	3.38	35.55	0.00	0.00	0.00	0.00	1.00	9095
# Non-Data Driven Investments	3.15	31.54	0.00	0.00	0.00	0.00	1.00	9095
# Data Investments	0.23	4.74	0.00	0.00	0.00	0.00	0.00	9095
SFR	876.14	3280.24	0.22	31.74	100.08	420.79	1797.87	9095
EQI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9095
RECPI	2.06	11.90	0.00	0.02	0.08	0.46	2.46	9095
GDP	4041297	12598352	72767	517446	1066089	2812575	7851039	9095
Income	36103	9321	22247	29821	34332	40337	47398	9095
Population	82867	216637	1729	13887	29595	67120	174788	9095

Table 2: **Data-Driven versus Traditional Summary Statistics**

This table reports the t-test of mean differences between Data-Driven VCs and Traditional (or non Data-Driven) VCs. In Panel A, all traditional VCs are included. In Panel B, only matched Traditional VCs are included. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

	Data-Driven		Traditional		Difference
	Mean	Count	Mean	Count	Data-Driven − Traditional
Panel A: VC-Year Level - All Traditional					
Age	14.65	598	11.92	7915	2.73***
# Employees	23.18	598	8.95	7915	15.17***
AUM ($ Mil)	1,256.73	598	437.90	7915	818.83***
Centrality	5.93	598	2.69	7915	3.23***
# Investments	22.47	598	7.85	7915	14.62***
# Hub Investments	16.84	598	5.02	7915	11.82***
# Non Hub Investments	5.46	598	2.80	7915	2.66***
# Low Comzone	0.47	598	0.27	7915	0.20***
Panel B: VC-Year Level - Matched Traditional					
Age	13.79	398	13.23	358	0.56
# Employees	23.83	398	20.31	358	3.52**
AUM ($ Mil)	1228.37	398	1162.34	358	66.02
Centrality	5.44	398	5.37	7915	0.07
# Investments	25.37	398	21.61	358	3.76**
# Hub Investments	19.02	398	16.83	358	2.21*
# Non Hub Investments	6.36	398	5.33	358	1.03**
# Low Comzone	0.5	398	0.39	358	0.11**

Table 3: **Data Technology Adoption and Number of Investments**

This table reports results for regressions at the VC-year level, investigating whether data-driven VCs make more investments per year after they adopt data technologies than traditional VCs. The dependent variable is the number of investments made by a VC firm in a given year. Panel A shows the number of total investments. Panel B shows the number of investments in Hubs. Panel C shows the number of investments in Low Activity areas. "State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression. Standard errors are clustered at the VC-firm level. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcomes:	# Investments							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: Total Investments				
Data Driven	0.790***	0.688***	0.113	0.026	0.224***	0.231***	0.144**	0.135**
	(6.14)	(5.98)	(1.27)	(0.46)	(2.93)	(3.14)	(2.26)	(2.06)
Log(VC Firm Age)		0.402***	-0.041	-0.163***		0.306***	-0.037	-0.308***
		(7.85)	(-1.11)	(-6.82)		(5.17)	(-0.61)	(-5.58)
Log(# Employees)			0.477***	0.207***			0.334***	0.213***
			(9.71)	(9.02)			(6.60)	(5.23)
Log(Total AUM)			0.223***	0.052***			0.190***	0.090***
			(6.18)	(3.25)			(5.97)	(3.58)
Centrality				0.190***				0.148***
				(26.15)				(16.95)
VC-Firm FE	No	No	No	No	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.34	0.38	0.52	0.63	0.66	0.66	0.67	0.68
N	8513	8513	8513	8513	8513	8513	8513	8513
				Panel B: Hub Investments				
Data Driven	0.635***	0.558***	0.016	-0.082	0.184	0.192*	0.098	0.101
	(5.08)	(4.73)	(0.17)	(-1.10)	(1.55)	(1.67)	(0.67)	(0.88)
Log(VC Firm Age)		0.322***	-0.061*	-0.169***		0.363***	-0.037	-0.303***
		(6.82)	(-1.66)	(-4.54)		(4.99)	(-0.54)	(-4.48)
Log(# Employees)			0.504***	0.322***			0.392***	0.279***
			(11.23)	(8.00)			(6.77)	(5.67)
Log(Total AUM)			0.155***	0.031			0.233***	0.135***
			(4.71)	(1.29)			(5.94)	(3.97)
Centrality				0.147***				0.154***
				(11.87)				(12.10)
VC-Firm FE	No	No	No	No	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.26	0.28	0.36	0.40	0.49	0.49	0.50	0.51
N	8513	8513	8513	8513	8513	8513	8513	8513

Panel C: Non Hub Investments

Data Driven	0.842***	0.730***	0.142	0.068	0.230***	0.237***	0.161**	0.152**
	(6.21)	(6.06)	(1.45)	(1.02)	(3.14)	(3.32)	(2.55)	(2.54)
Log(VC Firm Age)		0.434***	-0.036	-0.159***		0.282***	-0.040	-0.315***
		(7.53)	(-0.80)	(-5.79)		(4.53)	(-0.61)	(-5.15)
Log(# Employees)			0.463***	0.153***			0.310***	0.183***
			(8.02)	(6.28)			(5.58)	(4.11)
Log(Total AUM)			0.254***	0.061***			0.177***	0.076***
			(5.98)	(3.74)			(5.52)	(3.00)
Centrality				0.207***				0.147***
				(27.60)				(16.10)
VC-Firm FE	No	No	No	No	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.38	0.42	0.55	0.66	0.69	0.69	0.70	0.71
N	8513	8513	8513	8513	8513	8513	8513	8513

Panel D: Low Activity Investments

Data Driven	0.609***	0.547***	0.058	-0.010	0.481*	0.508**	0.442*	0.460*
	(3.66)	(3.55)	(0.40)	(-0.07)	(1.93)	(2.07)	(1.92)	(1.81)
Log(VC Firm Age)		0.294***	-0.044	-0.133**		0.228	-0.001	-0.259*
		(4.91)	(-0.78)	(-2.34)		(1.59)	(-0.00)	(-1.87)
Log(# Employees)			0.445***	0.303***			0.201*	0.038
			(7.59)	(5.06)			(1.78)	(0.35)
Log(Total AUM)			0.149***	0.031			0.245***	0.107
			(3.60)	(0.78)			(2.88)	(1.30)
Centrality				0.119***				0.174***
				(8.34)				(6.85)
VC-Firm FE	No	No	No	No	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.37	0.37	0.39	0.41	0.50	0.50	0.50	0.51
N	8513	8513	8513	8513	8513	8513	8513	8513

Table 4: **Data Technology Adoption and Growth**

This table reports results for regressions at the VC-year level. Panel A shows the impact of data technology adoption on total assets under management (columns (1) and (2)). Panel B shows the impact of data technology adoption on the number of partners (columns (1) and (2)) and the number of investments per partner (columns (3) and (4)). “State” denotes the state where the VC firm is headquartered. “Industry” denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). “Stage” denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). Standard errors are clustered at the VC-firm level. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Panel A: Fund Size

Outcomes:	Log(Total AUM)		Log(Median Round $)	
	(1)	(2)	(3)	(4)
Data Driven	0.481***	0.247**	0.100	0.006
	(2.75)	(2.39)	(0.95)	(0.08)
Log(VC Firm Age)	0.255***	0.397***	-0.033	0.032
	(4.28)	(5.22)	(-0.85)	(0.49)
Centrality	0.220***	0.120***	0.038***	0.013
	(13.73)	(7.20)	(4.47)	(1.44)
VC-Firm FE	No	Yes	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
R-squared	0.53	0.90	0.49	0.69
N	8513	8513	6075	6075

Panel B: Employee Size

	Log(# Partners)		Investment/Partner	
	(1)	(2)	(3)	(4)
Data Driven	0.209*	0.101	0.008	0.030
	(1.78)	(1.14)	(0.06)	(0.28)
Log(VC Firm Age)	0.227***	0.274***	-0.294***	-0.414***
	(6.17)	(6.22)	(-6.32)	(-5.58)
Centrality	0.071***	0.022**	0.156***	0.160***
	(7.05)	(2.55)	(13.13)	(8.96)
VC-Firm FE	No	Yes	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
R-squared	0.33	0.87	0.54	0.85
N	8513	8513	8513	8513

Table 5: **Data Technology Adoption and Number of Investments - Placebo**

This table reports results for regressions at the VC-year level, investigating whether data-driven VCs make more investments per year after they adopt data technologies than traditional VCs. The dependent variable is the number of investments made by a VC firm in a given year. Panel A shows the number of total investments. Panel B shows the number of investments in Hubs. Panel C shows the number of investments in Non Hubs. Panel D shows the number of investments in Low Activity areas. "State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression. Standard errors are clustered at the VC-firm level. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Treat =	Data Scientist Hire			Venture Partner Hire		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Total Investments						
Treat×Post	0.183**	0.183**	0.134*	0.168**	0.167**	0.113*
	(2.39)	(2.29)	(1.75)	(2.47)	(2.28)	(1.87)
Log(VC Firm Age)×Post	-0.158***	-0.158***	-0.192***	-0.157***	-0.158***	-0.193***
	(-5.01)	(-4.12)	(-5.04)	(-5.00)	(-4.14)	(-5.11)
Log(# Employees)×Post			0.045			0.051
			(1.13)			(1.28)
Log(Total AUM)×Post			0.038*			0.038*
			(1.76)			(1.73)
Centrality×Post		-0.000	-0.012		0.001	-0.012
		(-0.00)	(-1.18)		(0.06)	(-1.17)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.64	0.64	0.64	0.64	0.64	0.64
N	34293	33704	31069	34293	33704	31069
Panel B: Hub Investments						
Treat×Post	0.146*	0.119	0.104	0.184**	0.184**	0.116**
	(1.69)	(1.28)	(1.22)	(2.43)	(2.27)	(2.22)
Log(VC Firm Age)×Post	-0.171***	-0.187***	-0.197***	-0.171***	-0.188***	-0.198***
	(-4.25)	(-4.04)	(-4.29)	(-4.26)	(-4.05)	(-4.33)
Log(# Employees)×Post			-0.021			-0.018
			(-0.37)			(-0.33)
Log(Total AUM)×Post			0.033			0.033
			(1.25)			(1.24)
Centrality×Post		0.011	0.006		0.011	0.006
		(0.97)	(0.42)		(1.00)	(0.43)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.50	0.50	0.50	0.50	0.50	0.50
N	34293	33704	31069	34293	33704	31069

Panel C: Non Hub Investments						
Treat×Post	0.200**	0.201**	0.147*	0.090	0.088	0.082
	(2.44)	(2.36)	(1.67)	(0.93)	(0.90)	(0.84)
Log(VC Firm Age)×Post	-0.152***	-0.151***	-0.192***	-0.151***	-0.151***	-0.194***
	(-4.63)	(-3.79)	(-4.77)	(-4.60)	(-3.80)	(-4.85)
Log(# Employees)×Post			0.066*			0.073*
			(1.73)			(1.95)
Log(Total AUM)×Post			0.045**			0.044*
			(2.00)			(1.96)
Centrality×Post		-0.001	-0.016		-0.000	-0.016
		(-0.07)	(-1.63)		(-0.00)	(-1.62)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.68	0.68	0.68	0.68	0.68	0.68
N	34293	33704	31069	34293	33704	31069
Panel D: Low Activity Investments						
Treat×Post	0.502*	0.506*	0.511*	-0.291	-0.307	-0.071
	(1.81)	(1.80)	(1.77)	(-0.86)	(-0.88)	(-0.27)
Log(VC Firm Age)×Post	-0.113	-0.111	-0.106	-0.106	-0.114	-0.119
	(-1.06)	(-0.99)	(-0.97)	(-1.00)	(-1.02)	(-1.10)
Log(# Employees)×Post			-0.027			-0.013
			(-0.21)			(-0.11)
Log(Total AUM)×Post			0.013			0.022
			(0.17)			(0.29)
Centrality×Post		-0.001	-0.002		0.004	0.001
		(-0.06)	(-0.06)		(0.20)	(0.04)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.55	0.53	0.54	0.55	0.53	0.54
N	34293	33704	31069	34293	33704	31069

Table 6: **Data Technology Adoption and Number of Investments - IV Approach**

This table reports results for regressions for the instrumental variable two-stage least squares analysis at the VC-investment level, investigating whether the investments made by data-driven VCs after they adopt data technologies lead to different outcomes than those made by other VCs. Column (1) shows the first stage of the regression, where an indicator equal to one if an investment is made by a data-driven VC is fitted with the VC Exposure × New Fund. Columns (2) and (3) show investments made in non hub commuting zones, and columns (4) and (5) show investments made in low activity commuting zones. "State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression for columns (2) through (5). Standard errors are clustered at the VC-Firm level. For columns (3) and (5) standard errors are bootstrapped over 1,000 iterations.

Outcomes:		Non Hub CZ		Low Activity CZs	
	First Stage	PPML	2SLS	PPML	2SLS
	(1)	(2)	(3)	(4)	(5)
Data-Driven		0.170^{***}	0.872^{***}	0.688^{**}	1.436^{***}
		(3.33)	(2.45)	(2.75)	(2.55)
VC Exposure×New Fund	0.055^{***}				
	(3.71)				
Log(VC Firm Age)	-0.044	-0.332	-0.028	-0.021	-0.025
	(-0.78)	(-0.34)	(-0.54)	(-0.57)	(-0.57)
Log(# Employees)	0.445^{***}	0.303^{***}	0.281^{***}	0.285^{***}	0.264^{***}
	(7.59)	(5.06)	(4.78)	(5.35)	(5.23)
Log(Total AUM)	0.149^{***}	0.159^{***}	0.215^{***}	0.249^{***}	0.255^{***}
	(3.60)	(2.68)	(2.88)	(2.70)	(2.82)
Centrality	0.119^{***}	0.178^{***}	0.176^{***}	0.174^{***}	0.172^{***}
	(8.34)	(6.55)	(6.42)	(6.85)	(6.72)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes
F-Statistic	13.74				
R-squared		0.44	-0.05	0.45	-0.02
N	3301	3301	3301	3301	3301

Table 7: **Data Technology and Other Measures of Out-of-Network Investments**

This table reports results for regressions at the investment level, investigating the impact of data technology adoption on out-of-network investment outcomes.“State” denotes the state where the VC firm is headquartered. “Industry” denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). “Stage” denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcomes:	$\mathbb{1}$(Top Tercile Distance)		$\mathbb{1}$(Diff Industry)		$\mathbb{1}$(Local Syndicate)	
	(1)	(2)	(3)	(4)	(5)	(6)
Data-Driven	-0.004	0.027	0.089**	0.077*	-0.045	-0.020**
	(-0.13)	(1.33)	(2.15)	(1.80)	(-1.10)	(-2.03)
Log(VC Firm Age)	0.008	0.033***	0.020**	0.045**	-0.006	-0.013**
	(1.05)	(3.39)	(2.09)	(2.32)	(-1.18)	(-2.08)
Log(# Employees)	-0.015***	-0.022***	-0.017***	-0.026***	-0.018***	-0.026***
	(-2.71)	(-3.27)	(-2.75)	(-3.30)	(-2.70)	(-3.30)
Log(Total AUM)	0.008***	-0.000	0.008***	-0.000	0.008***	-0.000
	(3.62)	(-0.09)	(3.69)	(-0.09)	(3.62)	(-0.07)
Centrality	0.004***	0.001	0.005***	0.001	0.004***	0.002**
	(4.14)	(0.61)	(4.19)	(0.62)	(4.12)	(2.32)
Log(Startup Age)	0.010*	0.010**	-0.004	-0.003	-0.042***	-0.032***
	(1.89)	(2.20)	(-0.80)	(-1.04)	(-9.94)	(-8.97)
Serial	0.023**	0.021*	0.006	0.003*	-0.005	0.010
	(2.32)	(1.85)	(0.42)	(0.22)	(-0.20)	(0.64)
VC Prior	0.046*	0.045	-0.016	-0.015	-0.016	-0.014
	(1.85)	(1.97)	(-1.06)	(-0.97)	(-1.05)	(-0.96)
Alumni	0.016***	0.011***	0.018***	0.021***	0.016***	0.021***
	(2.99)	(3.42)	(3.21)	(3.66)	(2.84)	(3.23)
VC-Firm FE	No	Yes	No	Yes	No	Yes
Org-State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.15	0.24	0.22	0.37	0.45	0.50
N	49411	49411	49411	49411	49411	49411

Table 8: **Data Technology Adoption and Follow On Investments**

This table reports results for regressions at the investment level, investigating the impact of data technology adoption on investment outcomes. The dependent variable is an indicator equal to 1 if the startup receives another round of financing and 0 otherwise. Columns (1) through (3) include all investments and columns (4) through (6) include investments where the VC was the lead investor. "State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcome:	Follow On					
	All Investments			Lead Investor		
	(1)	(2)	(3)	(4)	(5)	(6)
Data Driven	0.026***	0.027***	0.029***	0.032**	0.036***	0.035***
	(3.40)	(3.34)	(3.75)	(2.41)	(2.81)	(2.70)
Data Driven×Non Hub		-0.004			-0.016	
		(-0.40)			(-0.76)	
Data Driven×Low Activity			-0.163***			-0.178**
			(-3.61)			(-2.31)
Non Hub		-0.002			-0.001	
		(-0.27)			(-0.15)	
Low Activity			-0.017			0.002
			(-1.05)			(0.09)
Log(VC Firm Age)	-0.025**	-0.025**	-0.025**	-0.043	-0.043	-0.043
	(-2.08)	(-2.09)	(-2.11)	(-1.57)	(-1.58)	(-1.58)
Log(# Employees)	-0.006	-0.006	-0.006	-0.015	-0.015	-0.016
	(-0.99)	(-0.99)	(-1.02)	(-0.97)	(-0.97)	(-0.99)
Log(Total AUM)	-0.001	-0.001	-0.001	0.006	0.006	0.006
	(-0.21)	(-0.21)	(-0.20)	(0.59)	(0.60)	(0.58)
Centrality	0.002	0.002	0.002	0.003	0.003	0.003
	(0.89)	(0.90)	(0.92)	(0.85)	(0.86)	(0.88)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.09	0.09	0.09	0.11	0.11	0.11
N	46871	46871	46871	14211	14211	14211
Data-Driven=Data-Driven×Non Hub (p-value)		0.0405**			0.0364**	
Non Hub=Data-Driven×Non Hub (p-value)		0.8487			0.5823	
Data-Driven=Data-Driven×Low Activity (p-value)			0.0001***			0.0049***
Low Activity=Data-Driven×Low Activity (p-value)			0.0023***			0.0412**

Table 9: **Data Technology Adoption and Exit through IPO or Achieve Unicorn Status**
This table reports results for regressions at the investment level, investigating the impact of data technology adoption on investment outcomes. The dependent variable is an indicator equal to 1 if the startup exits through an IPO or achieves unicorn status. Columns (1) through (3) include all investments and columns (4) through (6) include investments where the VC was the lead investor."State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcome:	Exit through IPO or Achieve Unicorn Status					
	All Investments			Lead Investor		
	(1)	(2)	(3)	(4)	(5)	(6)
Data Driven	0.013	0.007	0.012	-0.003	-0.014	-0.005
	(0.90)	(0.49)	(0.83)	(-0.14)	(-0.63)	(-0.20)
Data Driven×Non Hub		0.026**			0.041**	
		(2.25)			(2.36)	
Data Driven×Low Activity			0.045*			0.121*
			(1.87)			(1.72)
Non Hub		-0.019*			-0.021	
		(-1.79)			(-1.52)	
Low Activity			-0.054***			-0.045
			(-2.85)			(-1.47)
Log(VC Firm Age)	-0.030*	-0.029*	-0.030*	-0.057	-0.059	-0.058
	(-1.82)	(-1.78)	(-1.82)	(-1.42)	(-1.45)	(-1.43)
Log(# Employees)	0.008	0.008	0.008	0.041*	0.042*	0.041*
	(0.85)	(0.87)	(0.81)	(1.71)	(1.75)	(1.73)
Log(Total AUM)	0.002	0.002	0.002	0.004	0.005	0.004
	(0.29)	(0.29)	(0.29)	(0.27)	(0.31)	(0.26)
Centrality	-0.001	-0.001	-0.001	-0.006	-0.006	-0.006
	(-0.35)	(-0.38)	(-0.35)	(-0.95)	(-0.97)	(-0.95)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.16	0.16	0.16	0.08	0.08	0.08
N	22428	22428	22428	8526	8526	8526
Data-Driven=Data-Driven×Non Hub (p-value)		0.0245**			0.0099***	
Non Hub=Data-Driven×Non Hub (p-value)		0.0077***			0.0099**	
Data-Driven=Data-Driven×Low Activity (p-value)			0.0058***			0.0068***
Low Activity=Data-Driven×Low Activity (p-value)			0.1408			0.0401**

Table 10: **Data Technology Adoption and Exit through Acquisition**

This table reports results for regressions at the investment level, investigating the impact of data technology adoption on investment outcomes. The dependent variable is an indicator equal to 1 if the startup exits through an acquisition and 0 otherwise. Columns (1) through (3) include all investments and columns (4) through (6) include investments where the VC was the lead investor."State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcome:	Exit through Acquisition					
	All Investments			Lead Investor		
	(1)	(2)	(3)	(4)	(5)	(6)
Data Driven	0.002	0.007	0.002	0.016	0.016	0.017
	(0.09)	(0.29)	(0.08)	(0.58)	(0.61)	(0.61)
Data Driven×Non Hub		-0.022			-0.001	
		(-0.98)			(-0.04)	
Data Driven×Low Activity			0.013			-0.072
			(0.14)			(-0.35)
Non Hub		-0.018			-0.017	
		(-1.30)			(-0.90)	
Low Activity			-0.026			0.026
			(-0.89)			(0.54)
Log(VC Firm Age)	0.021	0.020	0.020	0.061	0.060	0.062
	(0.78)	(0.77)	(0.77)	(1.24)	(1.20)	(1.24)
Log(# Employees)	-0.011	-0.011	-0.011	-0.044	-0.044	-0.044
	(-0.72)	(-0.72)	(-0.74)	(-1.32)	(-1.30)	(-1.32)
Log(Total AUM)	0.004	0.004	0.004	-0.023	-0.022	-0.022
	(0.41)	(0.41)	(0.41)	(-1.18)	(-1.16)	(-1.17)
Centrality	0.000	0.000	0.000	0.007	0.007	0.007
	(0.06)	(0.06)	(0.06)	(1.02)	(1.03)	(1.02)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.14	0.14	0.14	0.13	0.13	0.13
N	22428	22428	22428	8526	8526	8526
Data-Driven=Data-Driven×Non Hub (p-value)		0.4333			0.6720	
Non Hub=Data-Driven×Non Hub (p-value)		0.8917			0.7288	
Data-Driven=Data-Driven×Low Activity (p-value)			0.9141			0.6679
Low Activity=Data-Driven×Low Activity (p-value)			0.7090			0.6520

Table 11: **Data Driven Investments in High Entrepreneurial Commuting Zones**

This table reports results for regressions at the commuting zone-year level, investigating which commuting zones attract more data-driven investments. The dependent variable is the number of investments made in a commuting zone in a given year. Panel A shows the number of investments made by data-driven VCs and Panel B the number of investments made by traditional VCs. Columns (1) and (2) show investments made in all commuting zones, columns (3) and (4) show investments made in non hub commuting zones, and columns (5) and (6) show investments made in low activity commuting zones. I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression. Standard errors are clustered at the commuting zone level. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcomes:	# Investments					
	All		Non Hub		Low Activity	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Data Driven						
Log(RECPI)	0.800***	-0.089***	0.725***	0.815***	0.421***	1.691*
	(5.40)	(-2.99)	(4.20)	(2.30)	(2.57)	(1.78)
Log(GDP)	0.616**	-0.277	0.338*	2.390	0.470	30.038**
	(2.54)	(-0.19)	(1.85)	(0.76)	(1.46)	(2.28)
Log(Income)	7.167***	-0.324	-3.216***	-5.455	-0.143	8.195
	(6.88)	(-0.21)	(-3.32)	(-1.45)	(-0.14)	(0.58)
Percent College	-3.565	-0.343	13.255***	8.007	10.886***	-142.184***
	(-1.18)	(-0.08)	(6.29)	(0.87)	(3.42)	(-2.98)
Comzone FE	No	Yes	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.89	0.95	0.60	0.76	0.23	0.79
N	5368	5368	5328	5328	4595	4595
Panel B: Non-Data Driven						
Log(RECPI)	0.792***	0.013	0.848***	0.006	0.395***	0.124
	(4.79)	(0.91)	(7.12)	(0.53)	(3.98)	(0.40)
Log(GDP)	0.385	-1.482	0.214	-0.314	0.880***	3.453*
	(1.50)	(-1.53)	(1.39)	(-0.36)	(5.24)	(1.70)
Log(Income)	3.546***	0.671	-2.417***	0.664	-0.545	0.029
	(3.12)	(0.47)	(-4.13)	(0.55)	(-1.01)	(0.01)
Percent College	2.615	2.630	11.489***	5.613	4.886***	2.293
	(0.85)	(0.78)	(6.74)	(1.59)	(3.26)	(0.18)
Comzone FE	No	Yes	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.90	0.97	0.78	0.91	0.33	0.61
N	5368	5368	5328	5328	4595	4595

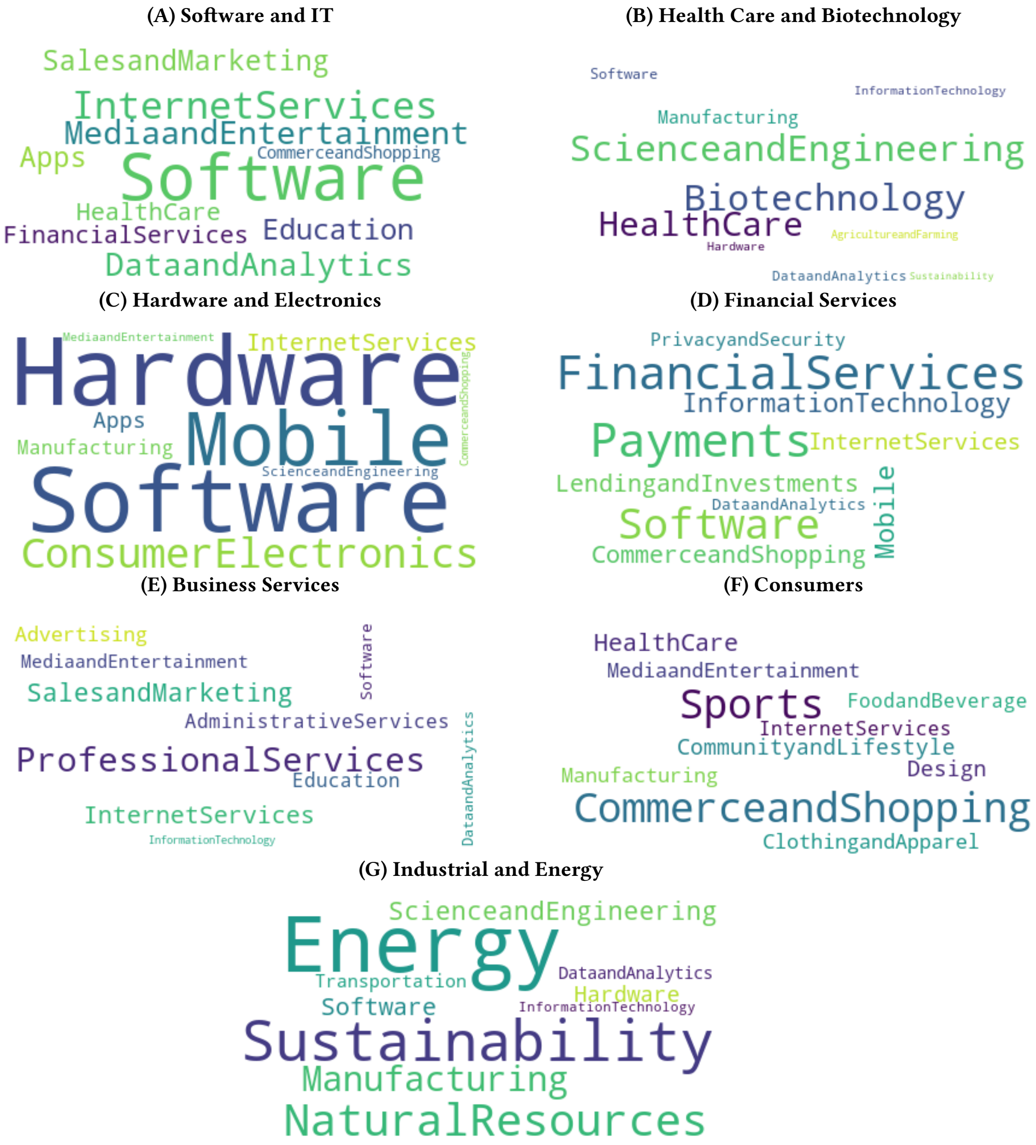
Table 12: **Data-Driven Investment Entry and VC Activity**

This table reports results for the stacked difference-in-difference regression at the county-level, investigating how VC activity changes in counties after entry of a data-driven VC. In columns (1) and (2), the dependent variable is the number of startups that receive their first ever funding rounds. In columns (3) and (4), the dependent variable is the number of patents produced by VC-backed startups. In columns (5) and (6), the dependent variable is the number of patents produced by entrepreneurial firms. Panel A includes all commuting zones with 25 or fewer VC investments in the previous five years. Panel B includes all commuting zones more than 1 but fewer than 25 VC investments in the previous 5 years. All columns include cohort by year fixed effects and cohort by commuting zone fixed effects. All columns include pre-data-entry VC activity controls. Even columns include pre-data-entry county level controls. Regressions are Poisson and are clustered at the commuting-zone level.

Outcome:	# Funding Rounds		# First VC Financing		# Unique Investors		# First Investor		# VC Patents	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treat×Post	0.116***	0.077**	0.120***	0.084**	0.256***	0.186***	0.178***	0.112**	0.347***	0.282***
	(2.81)	(2.59)	(2.93)	(2.63)	(3.55)	(3.12)	(2.82)	(2.44)	(3.83)	(3.64)
RECPI×Post		0.004		0.004		0.011**		0.011**		0.005
		(1.39)		(1.23)		(2.48)		(2.41)		(1.09)
Income×Post	0.168	0.147	0.027	0.001	0.350	0.207	0.229	0.080	0.590**	0.562**
	(0.84)	(0.71)	(0.11)	(0.00)	(1.25)	(0.72)	(0.73)	(0.25)	(2.51)	(2.40)
GDP×Post	0.006	-0.002	0.001	-0.008	-0.068*	-0.104***	-0.051	-0.087**	-0.027	-0.038
	(0.23)	(-0.07)	(0.03)	(-0.23)	(-1.92)	(-2.69)	(-1.29)	(-2.04)	(-0.86)	(-1.09)
Perc College×Post	0.399	0.479	0.305	0.397	1.628**	1.997**	2.266***	2.615***	-1.121*	-0.976*
	(0.68)	(0.80)	(0.44)	(0.56)	(2.09)	(2.50)	(2.62)	(2.95)	(-1.95)	(-1.66)
Cohort×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×Commuting Zone FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.75	0.75	0.65	0.64	0.76	0.76	0.72	0.72	0.68	0.68
N	53625	53548	53625	53548	53625	53548	53625	53548	53625	53548

Appendix

Figure A1: **Industry Classifications**



(A) Software and IT

(B) Health Care and Biotechnology

(C) Hardware and Electronics

(D) Financial Services

(E) Business Services

(F) Consumers

(G) Industrial and Energy

Table A1: **Data Technology Adoption and Number of Investments**

This table reports results for regressions at the VC-year level, investigating whether data-driven VCs make more investments per year after they adopt data technologies than traditional VCs. The dependent variable is the number of investments made by a VC firm in a given year. Panel A restricts the sample to VCs located in Hub areas. Panel B restricts the sample to first-time investments made by VC j in a startup. "State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression. Standard errors are clustered at the VC-firm level. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcomes:	# Total		# Hub		# Non Hub		# Low Activity	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Hub VCs								
Data Driven	0.002	0.165**	-0.055	0.118	0.028	0.175**	-0.020	0.444***
	(0.03)	(2.22)	(-0.64)	(0.95)	(0.39)	(2.50)	(-0.12)	(2.55)
Log(VC Firm Age)	-0.282***	-0.481***	-0.301***	-0.504***	-0.272***	-0.488***	-0.187***	-0.404**
	(-10.12)	(-6.98)	(-7.32)	(-6.69)	(-9.06)	(-6.61)	(-3.04)	(-2.45)
Log(# Employees)	0.230***	0.245***	0.354***	0.337***	0.176***	0.217***	0.330***	0.060
	(7.64)	(5.00)	(7.16)	(6.59)	(6.02)	(4.14)	(4.97)	(0.43)
Log(Total AUM)	0.009	0.042	-0.009	0.101***	0.016	0.029	-0.007	-0.062
	(0.47)	(1.28)	(-0.34)	(2.78)	(0.93)	(0.86)	(-0.17)	(-0.73)
Centrality	0.187***	0.140***	0.146***	0.136***	0.203***	0.142***	0.110***	0.172***
	(21.13)	(13.30)	(9.82)	(10.36)	(24.57)	(12.73)	(7.10)	(6.29)
VC-Firm FE	No	Yes	No	Yes	No	Yes	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.50	0.57	0.30	0.45	0.49	0.55	0.27	0.40
N	5960	5960	5960	8513	5960	5960	5960	5960
Panel B: First Time Investments								
Data Driven	-0.009	0.160**	-0.077	0.118	0.023	0.173**	-0.054	0.435***
	(-0.14)	(2.13)	(-0.89)	(0.95)	(0.31)	(2.46)	(-0.33)	(2.52)
Log(VC Firm Age)	-0.276***	-0.492***	-0.265***	-0.504***	-0.279***	-0.492***	-0.215***	-0.387**
	(-10.47)	(-7.43)	(-7.23)	(-6.69)	(-9.44)	(-6.76)	(-3.89)	(-2.43)
Log(# Employees)	0.241***	0.263***	0.361***	0.337***	0.180***	0.227***	0.357***	0.111
	(8.49)	(5.74)	(8.46)	(6.59)	(6.28)	(4.41)	(5.98)	(0.95)
Log(Total AUM)	0.013	0.045	-0.006	0.101***	0.022	0.030	-0.008	0.016
	(0.71)	(1.42)	(-0.22)	(2.78)	(1.26)	(0.88)	(-0.20)	(0.19)
Centrality	0.184***	0.140***	0.139***	0.136***	0.203***	0.142***	0.108***	0.167***
	(21.46)	(13.53)	(10.22)	(10.36)	(24.73)	(12.83)	(7.52)	(6.36)
VC-Firm FE	No	Yes	No	Yes	No	Yes	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.52	0.58	0.35	0.45	0.55	0.61	0.40	0.50
N	8513	8513	8513	8513	8513	8513	8513	8513

Table A2: **Data Technology Adoption and Number of Investments**

This table reports results for regressions at the VC-year level, investigating whether data-driven VCs make more investments per year after they adopt data technologies than traditional VCs. The dependent variable is the number of investments made by a VC firm in a given year. Panel A use the natural log of the number of data-related employees to proxy for Data Driven. Panel B uses the number of data-related employees scaled by number of partners to proxy for Data Driven. "State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression. Standard errors are clustered at the VC-firm level. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcomes:	# Total		# Hub		# Non Hub		# Low Activity	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: Log(# Data Employees)				
Data Driven	0.043	0.180***	0.013	0.171	0.057	0.177***	0.078	0.389**
	(1.34)	(3.33)	(0.19)	(1.60)	(1.56)	(3.67)	(0.82)	(2.09)
Log(VC Firm Age)	-0.162***	-0.299***	-0.167***	-0.290***	-0.158***	-0.308***	-0.131**	-0.254*
	(-6.77)	(-5.43)	(-4.47)	(-4.37)	(-5.78)	(-5.01)	(-2.30)	(-1.82)
Log(# Employees)	0.202***	0.210***	0.313***	0.273***	0.150***	0.182***	0.294***	0.045
	(8.61)	(5.34)	(7.65)	(5.53)	(5.95)	(4.21)	(4.86)	(0.41)
Log(Total AUM)	0.053***	0.089***	0.031	0.133***	0.063***	0.075***	0.032	0.103
	(3.26)	(3.55)	(1.25)	(3.91)	(3.81)	(2.96)	(0.80)	(1.24)
Centrality	0.190***	0.146***	0.147***	0.152***	0.207***	0.146***	0.118***	0.173***
	(25.63)	(16.66)	(11.71)	(11.91)	(27.16)	(15.83)	(8.25)	(6.82)
VC-Firm FE	No	Yes	No	Yes	No	Yes	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.63	0.68	0.40	0.51	0.66	0.71	0.41	0.51
N	8513	8513	8513	8513	8513	8513	8513	8513
				Panel B: Proportion of Data Employees				
Data Driven	0.163	0.707**	0.037	0.378	0.228	0.812***	-0.025	1.218**
	(0.98)	(2.49)	(0.12)	(0.74)	(1.14)	(3.06)	(-0.05)	(2.32)
Log(VC Firm Age)	-0.160***	-0.297***	-0.171***	-0.283***	-0.154***	-0.307***	-0.132**	-0.355**
	(-6.63)	(-5.00)	(-4.34)	(-3.86)	(-5.57)	(-4.71)	(-2.16)	(-2.47)
Log(# Employees)	0.201***	0.212***	0.312***	0.284***	0.152***	0.181***	0.302***	0.044
	(8.19)	(4.85)	(7.02)	(4.99)	(5.74)	(3.92)	(4.70)	(0.35)
Log(Total AUM)	0.055***	0.081***	0.034	0.127***	0.065***	0.067***	0.020	0.088
	(4.02)	(3.09)	(1.36)	(3.46)	(4.44)	(2.58)	(0.46)	(0.97)
Centrality	0.187***	0.146***	0.143***	0.150***	0.204***	0.146***	0.119***	0.178***
	(27.37)	(15.98)	(11.49)	(11.15)	(28.45)	(15.45)	(7.91)	(6.70)
VC-Firm FE	No	Yes	No	Yes	No	Yes	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.63	0.69	0.40	0.51	0.66	0.71	0.39	0.49
N	7326	7326	7326	7326	7326	7326	7326	7326

Table A3: **Data Technology Adoption and Number of Investments**

This table reports results for regressions at the VC-year level, investigating whether data-driven VCs make more investments per year after they adopt data technologies than traditional VCs. The dependent variable is the number of investments made by a VC firm in a given year. Panel A shows the number of total investments. Panel B shows the number of investments in Hubs. Panel C shows the number of investments in Non Hubs. Panel D shows the number of investments in Low Activity areas. "State" denotes the state where the VC firm is headquartered. "Industry" denotes the main industry the VC firm invests in over the sample period (among seven industries: Business Services, Consumers, Financial Services, Hardware and Electronics, Health Care and Biotechnology, Industrial and Energy, and Software and IT). "Stage" denotes the main stage the VC firm invests in over the sample period (among six categories: Pre-Seed, Seed, Series A, Series B, Series C, Series D and onward). I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression. Standard errors are clustered at the VC-firm level. The symbols $*$, $**$, and $***$ indicate significance at the 10%, 5%, and 1% level, respectively.

Outcomes:	Treated			Placebo		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Total Investments						
Treat×Post	0.206***	0.202***	0.173***			
	(4.94)	(4.79)	(4.09)			
Placebo×Post				0.088***	0.083***	0.060*
				(2.82)	(2.62)	(1.91)
Log(VC Firm Age)×Post	-0.169***	-0.172***	-0.194***	-0.167***	-0.171***	-0.194***
	(-16.44)	(-15.14)	(-16.11)	(-16.19)	(-14.97)	(-16.13)
Centrality×Post		0.002	-0.007**		0.002	-0.007**
		(0.79)	(-2.49)		(1.04)	(-2.48)
Log(# Employees)×Post			0.023*			0.030**
			(1.93)			(2.50)
Log(Total AUM)×Post			0.033***			0.032***
			(4.37)			(4.28)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×StageXYear FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.64	0.64	0.64	0.64	0.64	0.64
N	32781	32195	29690	32781	32195	29690
Panel B: Hub Investments						
Treat×Post	0.159**	0.123*	0.140**			
	(2.57)	(1.92)	(2.23)			
Placebo×Post				0.116**	0.087*	0.086*
				(2.27)	(1.69)	(1.66)
Log(VC Firm Age)×Post	-0.172***	-0.192***		-0.171***	-0.191***	-0.187***
	(-11.88)	(-12.31)		(-11.85)	(-12.27)	(-11.50)
Centrality×Post		0.014***	0.013***		0.014***	0.014***
		(4.04)	(3.44)		(4.08)	(3.21)
Log(# Employees)×Post			-0.041***			-0.049***
			(-2.64)			(-2.87)
Log(Total AUM)×Post						0.025**
						(2.52)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.50	0.50	0.50	0.50	0.50	0.50
N	32781	32195	31843	32781	32195	29690

Panel C: Non Hub Investments

Treat×Post	0.220***	0.218***	0.159***			
	(4.96)	(4.90)	(3.55)			
Placebo×Post				0.085***	0.082**	0.052
				(2.67)	(2.52)	(1.60)
Log(VC Firm Age)×Post	-0.167***	-0.168***		-0.164***	-0.166***	-0.199***
	(-14.70)	(-13.41)		(-14.38)	(-13.22)	(-14.90)
Centrality×Post		0.001	-0.014***		0.001	-0.012***
		(0.29)	(-4.76)		(0.56)	(-4.10)
Log(# Employees)×Post			0.055***			0.057***
			(4.38)			(4.43)
Log(Total AUM)×Post			0.039***			0.039***
			(4.70)			(4.60)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.68	0.68	0.69	0.68	0.68	0.69
N	32781	32195	29690	32781	32195	29690

Panel D: Low Activity Investments

Treat×Post	0.506**	0.500**	0.510**			
	(2.46)	(2.41)	(2.45)			
Placebo×Post				-0.061	-0.076	-0.090
				(-0.35)	(-0.44)	(-0.52)
Log(VC Firm Age)×Post	-0.104**	-0.108**	-0.097	-0.098*	-0.108**	-0.108*
	(-1.99)	(-1.97)	(-1.64)	(-1.87)	(-1.98)	(-1.81)
Centrality×Post		0.002	0.001		0.006	0.003
		(0.19)	(0.11)		(0.57)	(0.23)
Log(# Employees)×Post			-0.054			-0.040
			(-0.99)			(-0.73)
Log(Total AUM)×Post			0.025			0.033
			(0.74)			(0.96)
Cohort×VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.56	0.54	0.55	0.56	0.54	0.55
N	32781	32195	29690	32781	32195	29690

Table A4: **AI Glossary**

Artificial Intelligence (AI)
Artificial General Intelligence (AGI)
Algorithm
Anthropomorphism
Big Data
ChatGPT
Chatbot
Convolutional Neural Network (CNN)
Corpus
Copilot
Cutoff Date
Data Mining
Data Validation
Dall-E
Deepfake
Deep Learning
Embodied Agent
Expert System
Inception Distance
Intelligent Agent
Garbage in Garbage Out
Graphics Processing Unit (GPU)
Generative Pretrained Transformer (GPT)
Knowledge Engineering
Large Language Model
Machine Learning
Moats
Model Collapse
Natural Language Generation (NLG)
Natural Language Processing (NLP)
Neural Network
Neuromorphic Computing
OpenAI
Overfitting
Prompt Engineering
QLearning
Recommendation Engine
Reinforcement Learning
Sentiment Analysis
Supervised Learning
Speech Recognition
Synthetic Data
Technological Singularity
Transformer Model
Turing Test
Unsupervised Learning
Variational Autoencoder
Zeroshot Learning

Table A5: **Industry Exposure to AI**

Industry	Exposure	%Desc w/ Match	Nb. Desc w/ match	Nb. Descriptions
		Panel A: Most Exposed Industries		
Machine Learning	99	79.58	9921	12466
Artificial Intelligence	99	74.30	15975	21501
NLP	99	63.93	906	1417
Text Analytics	99	48.67	175	359
Speech Recognition	99	47.67	215	451
Computer Vision	99	45.42	824	1814
Facial Recognition	98	43.23	83	192
Predictive Analytics	98	39.68	988	2490
Data Mining	98	37.82	462	1171
Big Data	98	35.56	3523	9315
		Panel B: Least Exposed Industries		
Timber	0	0	0	362
Sailing	0	0	0	323
Comics	0	0	0	197
Bakery	0	0	0	1296
Wood Processing	0	0	2	2199
Theatre	1	0.1	1	1036
Laundry	1	0.1	1	969
Cosmetic Surgery	1	0.1	6	4464
Residential	1	0.15	39	22956
Winery	1	0.17	3	1668