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

Melissa Crumling*

Drexel University

September 5, 2025

Abstract

Information frictions make it difficult for venture capitalists (VCs) to identify promising startups outside established networks. I examine whether using data platforms and artificial intelligence tools that track and evaluate startups lowers these frictions and changes how VCs make investment decisions. After adoption, VCs increase investments outside their traditional networks by 13-35%, as proxied by startups located outside their geographic and industry focus. Plausibly exogenous variation in VCs' ability to hire data-related employees suggests these effects are causal. Conditional on surviving early funding rounds, these out-of-network investments are more likely to achieve major exits. At the regional level, entry by data-driven VCs into areas with historically low VC activity raises subsequent VC investment by 10-30%. These findings indicate that data technologies expand the reach of VCs while improving access to capital for founders beyond established networks.

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, Eliezer Fich, Laura Fields, Ioannis Floros (discussant), Geoffrey Friesen (ACE mentor), Tanja Kirmse, Alice Liu (discussant), Torin McFarland, Adriano Rampini, Chris Severen, Alex Tuft (discussant) and participants of the Rising Scholars Conference 2023, Philadelphia Fed Mentoring and Advancing PhDs program 2024, University of Delaware 2024, Financial Management Association (Doctoral Consortium) 2024, and WEFI (Workshop for Entrepreneurial Finance and Innovation) Fellows 2025 for helpful comments.

Acknowledgments

The author reports no financial support or conflicts of interest related to this project. Research access to Crunchbase data was provided through an approved API.

1 Introduction

Information frictions pose significant challenges in finance, particularly in the selection and screening of investment opportunities (Akerlof (1970); Stiglitz and Weiss (1981)). Firms face a tradeoff between protecting their competitive advantage and disclosing information to secure financing, which often leaves investors with limited insight into potential targets (Myers and Majluf (1984)). In recent years, however, new datasets and analytical tools have given investors alternative ways to evaluate investment opportunities (Zhu (2019)). Firms increasingly leave digital footprints through websites and other online platforms, while new datasets make it easier to track macroeconomic and industry trends. At the same time, advances in machine learning and artificial intelligence help synthesize this information at scale. These developments raise the following questions: to what extent do data technologies mitigate information frictions in financial markets, and how might this affect which firms ultimately receive financing?

To examine these questions, I focus on the venture capital (VC) industry, a setting where information frictions are particularly salient. Often considered the gatekeepers of high-growth startup financing (Gornall and Strebulaev (2021)), VCs invest in young, innovative firms with limited track records, which makes identifying and evaluating promising opportunities particularly challenging. Deal selection is central to VC performance (Sørensen (2007); Gompers et al. (2020)), which means generating high-quality deal flow is a first-order concern. To overcome information barriers, the industry has become highly concentrated, both geographically and within established investor and industry networks (Sorenson and Stuart (2001)). While such concentration can facilitate information flow (Hochberg, Ljungqvist, and Lu (2007)), it may also restrict access to capital for founders outside these networks, limiting the efficiency of capital allocation.[1]

In recent years, however, a growing number of VCs have adopted data technologies to support the pre-investment process, including sourcing, evaluating, and selecting startups.[2]. Firms are increasingly building their own proprietary platforms to track and analyze hundreds of startups

1. For instance, see WSJ: Why It Pays for Startups to Move to Silicon Valley and Medium: Ban Warm Introductions.

2. According to Data-Driven VC Landscape 2025, 79% of firms that use data technologies do so for better deal coverage and better screening: DDVC Landscape 2025. See also FT: Are machines smarter than venture capitalists? and WSJ: VC Firms Have Long Backed AI. Now, They Are Using It.

simultaneously, using machine learning tools to extract signals from alternative data sources.[3] These developments represent a shift from relationship-driven decision making to more data-augmented screening.

In the first part of the paper, I examine whether data technologies expand VCs' investment opportunity set. The overarching prediction is that adopting these tools reduces search frictions, allowing VCs to identify promising startups beyond their established networks. To measure adoption, I use detailed employee and job description data from Crunchbase and LinkedIn. I classify a VC as data-driven starting from the year of their first data scientist hire. This hiring serves as a credible indicator of internal adoption, as data platforms typically require high-skilled human capital to build, operate, and interpret. [4] Human capital has been used as a proxy for technology adoption in prior work in VC (Retterath (2020); Bonelli (2025)) and in other financial markets (e.g. Raymond (2023); Babina et al. (2024)). As shown in Figure 1, adoption of data technologies among VC firms has been rising steadily over the past decade.

To examine whether data technologies shift investment beyond established networks, I begin with the geographic distribution of VC activity in the United States (US). More than two-thirds of VC investments are concentrated in three major hubs – California (CA), Massachusetts (MA), and New York (NY).[5] Geographic proximity helps mitigate information frictions, leading investors to headquarter and invest locally in these regions (Sorenson and Stuart (2001); Cumming and Dai (2010)). VCs are also more likely to establish satellite offices in these hubs (Chen et al. (2010)), reinforcing network-based interactions across geographic markets. Consequently, startups outside CA, MA, and NY ("non-hub" regions) are more likely to be excluded from traditional investor networks. To provide a conservative test, I focus on VCs headquartered in these three states and examine whether adopting data technologies increases their investments in non-hub locations.

The first set of tests examines whether data technologies influence which deals are selected (i.e. the extensive margin). Because only completed investments are observable, I construct a set of counterfactual investments from both the startup and investor perspectives.[6] For each investor

3. For greater detail see Medium: The Future of VC: Augmenting Humans with AI.

4. Data Driven VC, a website that tracks usage of these tools by VCs, classifies investors as Data Driven if they have one or more engineers on the investment team.

5. NVCA 2020: 85% of capital raised and 73% of capital invested by VCs and startups are located in CA, MA, and NY.

6. See Hegde and Tumlinson (2014), Garfinkel et al. (2025), and Venugopal and Yerramilli (2025) who use similar

in an actual deal, I identify four control startups the investor could plausibly have considered but did not fund. These control startups are selected to operate in the same industry, be funded in the same year and stage (e.g., seed, Series C), and involve investors located in the same state as the actual investor. Likewise, for each funded startup, I identify four counterfactual investors located in the same state as the actual investor who invested in another firm in the same industry-stage in the same year.

Using this sample of actual and counterfactual matches, I find that data-driven VCs are more likely than traditional VCs to invest in non-hub startups. Adoption of data technologies increases the probability of investing in a non-hub startup by 0.46-0.53 percentage points, a 4% increase relative to the 13% average investment probability. Further tests indicate that the effect is stronger when the non-hub startup is more likely to fall outside of *other* professional networks. Specifically, I find that data-driven investments in non-hub startups are more likely when the startup is located in a region without an active VC market, when the founder attended a university outside the U.S. top 20 universities, and when the founder is a first-time entrepreneur.

Next, I examine whether data technology adoption increases the number of investments in non-hub startups (i.e., the intensive margin). Focusing on VCs headquartered in hub states, I find that after adopting data technologies, they increase the number of investments in non-hub startups by 20-35% per year (≈ 0.5 additional deals annually), compared to other VCs in the same state-industry-stage-year segment. Importantly, this growth is not merely proportional to their overall activity: data-driven VCs tilt their portfolios toward non-hub startups, raising the share of non-hub investments by 14% relative to hub investments.

Beyond geography, VCs also tend to specialize in particular industries (Hochberg, Mazzeo, and McDevitt (2015)), which can limit exposure to novel opportunities. Because industry specialization is less binary than office location, I use a VC-year panel and focus on intensive-margin tests. In the full sample, data-driven VCs invest in 21% more industries than non-data-driven VCs in the same state-industry-stage segment. This broader reach is not purely a cross-sectional difference. Following adoption, data-driven VCs increase the number of investments outside their main industry by 13%. Given that the majority of data-driven VCs have a technology focus (Bonelli (2025)), I expect the largest shifts to occur in industries where they are less represented. Con-

approaches.

sistent with this, I find that following adoption, data-driven VCs increase their investments in healthcare and biotechnology startups by 32% and in energy and manufacturing by 18%, sectors in which they have historically been less active.

Taken together, these results show that data technology adoption is associated with greater investment in startups typically outside VC networks, as proxied by VCs' geographic and industry focus. To mitigate concerns about correlated unobservables, I include VC headquarter state $\times$ industry $\times$ funding-round stage $\times$ investment-year fixed effects, which absorb local time trends that may otherwise confound shifts away from a VC's customary geo-industry focus with data technology adoption. To address concerns that VCs using data technologies may differ from those that do not, I include VC firm fixed effects in the VC-panel tests, allowing for a comparison of VC investment decisions before and after technology adoption. Finally, to address selection, I implement an instrumental-variables strategy that leverages two sources of plausibly exogenous variation: pre-adoption exposure to AI and the timing of raising a new fund. The results remain robust to the IV approach, providing evidence of a causal relationship between data technology adoption and investment decisions.

In the second part of the paper, I examine how data-driven investments perform when they occur outside a VC's main network. Prior work finds that adopting data technologies helps identify startups less likely to fail but does not confer additional advantages for major exits (IPO or acquisition; Bonelli (2025)). Evidence is limited, however, on how data-driven investments perform outside an investor's established network. I find that out-of-network investments are less likely to receive follow-on financing, but conditional on receiving it, they are more likely to achieve a high-quality exit.

Further analysis indicates that these patterns are driven by two contemporaneous effects: less intense monitoring and improved deal selection. On the one hand, monitoring is more difficult when investing in a startup outside a VC's main focus. For example, non-hub startups are likely to be further from a VC's headquarters, making regular face-to-face interactions more challenging. As a result, VCs are more likely to abandon startups of similar quality in these settings compared to their local investments (Sorenson and Stuart (2001)). On the other hand, lower competition outside core networks may provide access to higher-quality opportunities, improving investment performance (Hochberg, Mazzeo, and McDevitt (2015)). Taken together, these find-

ings suggest that data technologies expand VCs' reach beyond traditional hubs to higher-quality deals. However, the ultimate success of such investments depends on overcoming weaker monitoring intensity.

In the final part of the paper, I examine whether certain regions benefit from the entry of data-driven VCs. If VCs increase investments in non-hub startups, particularly in regions without an active VC market ("low-activity areas"), these regions can experience an increase in subsequent VC activity (Chen and Ewens (2021)). There are several channels through which such effects occur. First, once a data-driven VC invests in a region, its own search and monitoring costs for evaluating additional startups in that area declines (Chen et al. (2010)). Second, information spillovers about potential opportunities arise through syndicate and professional networks (Sorenson and Stuart (2001), Hochberg, Ljungqvist, and Lu (2010)), which can attract new investors. In both cases, the first data-driven investment links the region to larger VC networks and raises its visibility among investors.

To test this, I construct a panel of all low-activity commuting zones in the United States during the sample period (defined as those with 25 or fewer VC investments over the prior five years; Hochberg, Ljungqvist, and Lu (2010)). I then identify commuting zones that receive their first data-driven investment and classify them as treated (56 zones), with all other low-activity zones serving as controls. I estimate a stacked difference-in-differences design, comparing outcomes before and after the first data-driven entry with outcomes in control zones over the same horizon. Outcomes include the number of funding rounds, the number of first-time VC financings, the number of distinct investors, and the number of VC-backed patents. Following data-driven entry, subsequent VC activity increases by 10–30% per year across these measures. Overall, these results indicate that entry by data-driven VCs has persistent effects on local entrepreneurial finance and innovation and suggest that data technologies can lower barriers to high-growth startup financing.

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 spillover effects of VC activity when data-

driven VCs invest in non-hub locations. 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. (2025), 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 contributes to the growing literature on the role of data technologies in the venture capital industry. Prior work has examined how the internet facilitates investment sourcing (e.g., Li, Li, and Yang (2022)), while more recent research explores the potential of artificial intelligence and machine learning. For example, Lyonnet and Stern (2022) and Davenport (2022) demonstrate that algorithms can outperform human investors in identifying promising startups ex ante, partly because VCs often rely on stereotypical heuristics. Similarly, Retterath (2020) develops a predictive model that more accurately identifies high-performing investments than actual VC decisions. The only other paper I am aware of that studies the ex-post impact of VCs adopting this technology is Bonelli (2025), which shows that data tools help VCs identify startups similar to those they have funded before but are less effective at uncovering novel opportunities. Overall, these studies primarily examine the predictive or screening capabilities of algorithms. In contrast, I focus on the discovery margin: how data technologies expand the set of firms that VCs consider, particularly outside of major VC hubs. I show that these broader, often overlooked investments can lead to high-quality outcome.

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 (D'Acunto, 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

3.1 Traditional VC Model

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 such as monitoring and advising startups (see Panel A Figure C2). VC returns generally stem from two main channels: (1) selecting high-quality deals and (2) post-investment support through monitoring, advising, and strategic guidance. Prior research emphasizes that the first channel, deal selection, accounts for the larger share of fund performance (Sørensen (2007); Gompers et al. (2020)). The ability to identify "home run" investments ex ante is viewed as a key determinant of success, with value-add activities providing complementary but comparatively smaller contributions.

The scale of deal sourcing highlights the challenge of selection. VCs evaluate a large number of potential startups each year, but ultimately invest in only a fraction. These targets are often early-stage firms, with business models that are not yet proven and in many cases pre-revenue. Passion Capital's first fund provides an example: over its first two years, the firm considered 3,464 startups (41% via website submissions, 28% via mutual contacts, and 31% via events) and invested in 34. Of the realized deals, 77% originated from personal referrals and 17% from events (see Panel B Figure C2). Survey evidence from Gompers et al. (2020) shows similar patterns: for every investment ultimately made, a VC considers roughly 100 potential opportunities, with about 60% of deals sourced through professional networks and referrals from other investors or portfolio companies. This illustrate the central role of networks in channeling opportunities and facilitating information flows.

Networks can take different forms. Geographic proximity and industry specialization shape how information is transmitted and shared across investors (Sorenson and Stuart (2001); Hochberg, Ljungqvist, and Lu (2007)). Educational ties also matter: for example, Garfinkel et al. (2025) show that a substantial share of deals occur between VCs and founders who attended the same alma mater. These patterns indicate that network-based connections are a primary mechanism by which VCs identify investment opportunities. In contrast, cold approaches such as website sub-

missions rarely result in completed deals, reflecting either differences in quality or the importance of endorsement through existing relationships. As a result, startups that are outside a VC's established network, or that lack a referral pathway, may have lower visibility in the deal sourcing process.

3.2 Data-Driven Model

Since deal selection is the primary driver of VC performance, investors seek to cast a wide net over potential startups to increase the odds of identifying "home run" opportunities (Sørensen (2007); Gompers et al. (2020)). In practice, however, most deal flow arrives via inbound channels, either through unsolicited business plan submissions or, more importantly, through professional networks. Prior literature shows that investors with larger networks tend to outperform (Hochberg, Ljungqvist, and Lu (2007)), but network-based sourcing does not guarantee comprehensive coverage of opportunities. Because investors cannot manually evaluate the vast number of potential startups each year, networks play a central role in filtering opportunities and reducing information frictions (Gompers et al. (2020)). As Damian Cristian, founder of Koble, notes, "*Above all else, VC is a network business, effectively capped by the scalability of human relationships.*"

Data-driven approaches expand coverage by systematizing the early stages of the investment process. Early adopters hire data scientists to build internal data pipelines and tools. Examples include SignalFire's BeaconAI, Tribe Capital's Termina, and Connetic Ventures' Wendel (Panel A, Figure C1). In broad terms, these platforms do four things[7]:

1. **Identify:** Continuously surface new firms from standard databases (e.g., PitchBook, Crunchbase) and non-obvious sources such as LinkedIn team changes, GitHub repositories, product-launch sites, news feeds, app stores, and new business registrations.

2. **Enrich:** Assemble a consolidated "company profile" by gathering information on the firm (team, location, industry) and "growth metrics" such as hiring velocity, web or app traffic, and social activity, often via APIs and web crawling.

7. See https://medium.com/birds-view/the-future-of-vc-augmenting-humans-with-ai-30f1d79a09c3 for more in-depth description.

3. **Score and rank:** Translate collected data into signals using deterministic scorecards and supervised ML models (e.g., gradient-boosted trees) to prioritize which startups should receive further review.

4. **Outreach:** Conduct proactive outreach when signals cross thresholds, shifting deal flow from inbound to outbound.

This approach allows VCs to evaluate far more startups than would be possible through manual methods alone. For example, SignalFire reports that BeaconAI tracks and ranks more than 80 million companies, 600 million people, and millions of open-source projects.[8] Basis Set Ventures describes using large language models to analyze founders' cognitive and behavioral traits as part of its evaluation process.[9] These capabilities expand the top of the funnel and standardize screening, enabling VCs to observe a broader range of potential opportunities. While data-driven techniques have so far been concentrated in sourcing and screening, they are increasingly being extended to later stages of the investment process such as portfolio management and fundraising (Panel B, Figure C1).

At the same time, human judgment remains a critical element of the VC process. Data platforms can scale the discovery and initial evaluation of startups, but assessing the novelty and long-term potential of young firms is not a strength of purely algorithmic approaches (Bonelli (2025)). Prior work shows that machine learning-based screening can perform at least as well as top investors when evaluated against historical labels, but these models inevitably mirror past patterns and therefore benefit from human checkpoints to avoid pattern lock-in and fairness concerns (Retterath (2020)). In practice, data-driven sourcing and screening complement rather than replace investors' expertise: platforms expand identification and standardize early evaluation, while investors focus their attention on diligence, access, and winning competitive deals. 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.*"

8. https://www.signalfire.com/blog/signalfire-beacon-ai
9. https://www.basisset.com/founder-superpowers

4 Data and Key Variables

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 [10]. I am left with 2,772 distinct VC firms during my sample period of 2010 to 2022 [11]. 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 2,772 unique investors, 26,180 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, headquarter 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. (2025), Huang (2022), and 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. Retterath (2020), Raymond (2023)), Bonelli (2025)), 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

10. 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.

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

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[12], 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[13]. Additionally, a 2018 Pitchbook survey found that 85% of VCs use data for sourcing investments, with 38% using data for all investment sourcing[14]. 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 65 matches. The remainder of the firms were 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 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 19 data-driven VCs. Following prior literature, I

12. https://www.datadrivenvc.io/
13. https://landscape2024.datadrivenvc.io/
14. https://pitchbook.com/media/press-releases/pitchbook-survey-finds-only-38-of-venture-capital-investors-currently-use-data-t

classify a VC as data-driven based on the hire date of its first data-related employee, resulting in 83 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 slightly older (15.29) than traditional VCs (13.88) and are also larger based on number of employees (10.72 versus 6.20) and assets under management (1.1 billion versus 0.1 billion). They also tend to invest in more startups per year (26.48 versus 4.90).

4.3 Controls

I include various VC- and startup-level controls, depending on the specification. VC controls consist of firm age (to proxy for experience), number of employees, and total assets under management (to capture organizational size). Because fund-level information is incomplete, I approximate assets under management as the total capital deployed by a VC over the previous five years.[15]

Startup controls include firm age (to proxy for lifecycle stage) and an indicator for whether the startup participated in an accelerator (to capture early-stage support and visibility). For extensive-margin tests that use counterfactual deals as controls, I also include a measure of whether the startup's founder or CEO attended the same university as the VC. Recent research shows that alumni connections form important networks that influence deal selection (Garfinkel et al. (2025); Huang (2022); Koenig (2022)).

4.4 Startup Performance

The first measure of startup performance is whether a startup receives a follow-on round of financing. This occurs in 61% of investment cases and serves as a proxy for intermediate success, since it indicates continued investor confidence and the ability to raise additional capital. By contrast, VCs ultimately earn the majority of their returns through exit events, primarily IPOs and high-quality acquisitions. These exits allow VCs to capitalize on the growth and value of

15. I calculate this by summing funding round sizes. If multiple investors participate in a round, I divide the amount equally across all investors.

their portfolio companies, translating their equity stakes into liquidity. Among exits, IPOs are generally regarded as the most lucrative outcome, often generating a 5x to 10x return and sometimes exceeding 20x, whereas acquisitions typically yield a more modest 1x to 5x return. Over the past decade, however, firms have tended to stay private longer (Ewens, Nanda, and Rhodes-Kropf (2018)) and often go public at a much older and larger stage (Gao, Ritter, and Zhu (2013)). Many of the most successful firms reach unicorn status (a valuation above $1 billion) before eventually pursuing an IPO.

I define a high-quality acquisition as one in which the purchase price exceeds the total amount of financing raised by the startup. In my sample, 23% of startups exit through a high-quality acquisition, 7% exit through an IPO, and 10% achieve unicorn status. While follow-on funding reflects a milestone in a startup's development, major exits represent the ultimate realization of VC returns. VC performance is therefore disproportionately driven by the right tail of the return distribution.

4.5 Industry Classifications

Crunchbase provides detailed industry classifications, comprising 49 industry groups and more than 700 granular industry categories.[16] Industry groups are broad domains that encompass multiple industries, whereas industries are specific market segments. Company profiles may be assigned to multiple industries and industry groups. Crunchbase encourages profiles to list three to five industries to improve discoverability, and these fields can be edited by verified company employees and by the broader Crunchbase contributor community within a taxonomy maintained by Crunchbase.[17]

Because there is no standard industry classification for startups, I construct a venture-focused classification using latent Dirichlet allocation (LDA) topic modeling. I estimate a 12-topic model on the set of Crunchbase firms that have ever received venture financing and map Crunchbase's 49 industry groups into these 12 consolidated categories. Although 12 groups are fewer than the number of categories in public-firm classifications such as SIC or the Fama and French 48 industries, this level of aggregation is reasonable within the venture literature. VentureXpert, for

16. Crunchbase: How are Industries Organized?
17. Crunchbase: Best Practices to Follow.

example, reports six major industry groups and ten minor groups. The model yields 12 groups, which I label Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software. Word clouds for each group are shown in Figure C3.

When aggregating to the VC year level, I classify each VC's main industry focus as the industry in which it made the most investments in the previous year. In the event of a tie, I pick an industry at random. For the first year a VC appears in the sample, I use the industry with the most investments in that year, with ties again broken at random.

5 Data-Driven Investors and Investment Opportunities

In this section, I examine whether data technologies reduce search frictions in identifying investment opportunities. Venture capital investments are inherently characterized by high levels of information asymmetry Dessein (2005). As a result, VCs often favor local investments Cumming and Dai (2010) or specialize within particular industries Hochberg, Mazzeo, and McDevitt (2015), as these boundaries can facilitate information flow. At the same time, such distinctions may limit the scope of the investment opportunity set, as VCs may overlook startups outside their established networks. Since data technologies can systematically identify startups with a digital presence, this section investigates whether these tools help VCs overcome traditional network constraints in sourcing investments. α_g are group fixed effects so the actual investments are compared to their counterfactuals. VC firm fixed effects to control for any time invariant VC characteristics.

5.1 Extensive Margin of VC Investment in Non Hub Areas

First, I test whether data technologies influence deal selection, that is, the extensive margin of venture funding in non-hub commuting zones. I focus on investors in hub areas, specifically in San Jose, CA; San Francisco, CA; Boston, MA; and New York, NY. This analysis is typically challenging because researchers observe only realized investments, not the full set of startups considered by investors. To address this, I follow prior studies (Hegde and Tumlinson (2014) and

Garfinkel et al. (2025)) and employ a conditional linear probability model using counterfactual investments. Specifically, for each observed startup-investor match in my sample, I generate eight counterfactual startup-investor pairs that are similar but did not result in an actual investment. The first four pairs are constructed by holding the startup fixed and pairing it with other investors located in the same state as the actual investor, who were active in the same year and invested in startups in the same industry and stage. This design allows me to test whether the adoption of data technologies explains why the actual investor selected the non-hub startup. The second four pairs are constructed by holding the investor fixed and pairing it with alternative startups that it did not fund, but that were funded in the same year by investors located in the same state and active in the same industry and stage.

For example, consider Lightspeed Venture Partners, a San Francisco-based VC in my sample that became data-driven in 2018. In 2022, Lightspeed invested in Fonbnk, a fintech startup located in Dover, DE, during its seed round. To construct the counterfactuals, I first select four other fintech startups that also raised seed rounds in 2022 but did not receive financing from Lightspeed; each of these startups was instead funded by another San Francisco-based VC active in fintech at the seed stage in 2022. Next, to identify the counterfactual investors, I select four other San Francisco-based VCs that did not invest in Fonbnk in 2022 but did participate in seed financings of other fintech startups that year.

The extensive margin test is implemented as a linear probability model where the dependent variable equals one if the deal is realized. The key regressor is the interaction between $DataDriven$, an indicator equal to one if the investor has adopted data technologies, and $NonHub$, an indicator equal to one if the startup is located outside of the main hub areas. This specification evaluates whether data-driven investors are more likely to form matches with non-hub startups. The regression is formally expressed as:

$$Deal_{i,j,t} = \beta_1 DataDriven_{j,t} \times NonHub_i + X_{i,j,t} + \alpha_g + \epsilon_{i,j,t} \tag{1}$$

The dependent variable, $Deal_{i,j,t}$, is an indicator equal to one if the deal is realized and zero otherwise. The main explanatory variable, $DataDriven_{j,t} \times NonHub_i$, is the interaction between whether the VC is data-driven and whether the startup is located in a non-hub commuting

zone. $X_{i,j,t}$ includes time-varying controls for VC j and startup i, such as VC age (controlling for experience), number of employees and total assets under management (controlling for organizational size), startup age (controlling for lifecycle stage), whether the startup participated in an accelerator prior to VC financing (controlling for early-stage support and visibility), and whether the startup's founder or CEO attended the same university as the VC (controlling for alumni connections). Group fixed effects, α_g, are included so that each actual deal is compared directly to its set of counterfactual startup-investor pairs. The coefficient β_1 is therefore identified by within-group comparisons and captures whether the adoption of data technologies increases the probability that VCs invest in non-hub startups relative to their counterfactual matches.

5.1.1 Main Results

Table 3 presents the results. Column (1) shows that the coefficient on $Non\ Hub$ is negative and statistically significant, indicating that startups in non-hub commuting zones are substantially less likely to receive VC investment from hub-based investors. In contrast, the interaction between $DataDriven$ and $Non\ Hub$ is positive and significant, increasing the likelihood of investment by 0.47 percentage points. Relative to the mean probability of investment of 13%, this represents a 3.5% increase in the likelihood of a data-driven VC investing in a non-hub startup. The regression controls for both investor and startup characteristics, including investor age and size (column 2), startup age and visibility (column 3), and alumni connections (column 4). In addition, I include group fixed effects, allowing each actual deal to be compared directly to its set of counterfactual matches. The coefficients remain stable across specifications, highlighting the persistent effect of data-driven investors being more likely to invest in non-hub startups.

5.1.2 Cross Sectional Heterogeneity

Next, I examine cross-sectional heterogeneity in the effect of data-driven investments in non-hub startups. I expect the effect to be stronger in two settings: when there is a lack of local VC financing and when startups are more likely, beyond location, to fall outside of traditional networks. I begin with the former. Because VC investing is a two-sided matching process Sørensen (2007), startups located in non-hub areas with active local VC markets may be less likely to accept offers from hub-based investors, as local VCs may provide more relevant support and monitoring Chen

et al. (2010). While roughly two-thirds of VC activity occurs in the three major hubs, there is a growing presence of venture activity in other regions. Based on both anecdotal evidence and patterns in the data, I identify the next eleven largest VC markets: Los Angeles, CA; Seattle, WA; Austin, TX; Chicago, IL; Denver, CO; San Diego, CA; Washington, DC; Atlanta, GA; Portland, OR; Miami, FL; and Philadelphia, PA. [18] Startups in these secondary hubs may benefit from the presence of a local VC ecosystem and thus be less reliant on hub-based, data-driven investors. To test this idea, I split $Non\ Hub$ into two groups and interact each with $DataDriven$: $Secondary\ Hub$, which includes the eleven regions above, and $Low\ Hub$, which includes all other non-hub commuting zones.

The results are presented in Columns (1) and (2) of Table 4. The interactions of $DataDriven \times Low\ Hub$ and $DataDriven \times Secondary\ Hub$ are both positive, but the coefficient on $DataDriven \times Low\ Hub$ is larger in magnitude (0.74 versus 0.42 in column (2)) and marginally more significant. This pattern suggests that data technologies are particularly valuable in regions with limited local financing options and higher geographic frictions, rather than in secondary hubs with more established venture markets.

Next, I examine whether data-driven VCs are more likely to invest in non-hub startups whose founders fall outside of traditional professional networks. Alumni ties and educational background are important components of professional networks Garfinkel et al. (2025). Founders who did not attend a top 20 university are therefore more likely to be outside of these networks.[19] My prediction is that data-driven investors will have a larger effect in funding non-hub startups with founders from outside top 20 universities. The results are displayed in Columns (3) and (4) of Table 4. The coefficient on $DataDriven \times NonHub \times NonTop20$ is positive and statistically significant at the 10% level, with magnitudes of 1.034 and 0.985. These estimates imply that non-hub startups with founders outside of top 20 universities are more likely to receive investment from data-driven VCs compared to otherwise similar non-hub startups with top 20 founders.

My final proxy for founders who are likely to fall outside of established professional networks

18. See WSJ: Colorado Emerges as a Venture-Capital Hub.

19. Top 20 universities according to U.S. News' 2024 rankings: Princeton University, Harvard University, Columbia University, Massachusetts Institute of Technology, Yale University, Stanford University, University of Chicago, University of Pennsylvania, Northwestern University, Duke University, Johns Hopkins University, California Institute of Technology, Dartmouth College, Brown University, University of Notre Dame, Vanderbilt University, Cornell University, Rice University, Washington University in St. Louis, and the University of California, Los Angeles (UCLA).

are first-time founders. Prior work shows that serial entrepreneurs are more likely to attract funding at earlier stages and to receive backing from more experienced VCs in their subsequent ventures (P. Gompers et al. (2010)). By contrast, first-time founders, particularly those located in non-hub regions, may face greater challenges accessing capital and thus stand to benefit more from data technologies. To test this idea, I interact $DataDriven \times NonHub$ with an indicator for $First\text{-}Time\ Founder$. The results, reported in Columns (5) and (6), show that while the coefficients are not statistically significant, the estimated effects are positive and the magnitudes are larger for $DataDriven \times NonHub \times First\text{-}Time\ Founder$ than for $DataDriven \times NonHub$. This pattern suggests that first-time founders in non-hub regions are relatively more likely to receive investment from data-driven VCs compared to their serial counterparts, even though the estimates remain imprecise.

Taken together, these results suggest that data technologies play their largest role in reducing information frictions in settings where alternative sources of capital and established networks are weakest, particularly in low hub regions, among founders from nonelite universities, and among first-time entrepreneurs.

5.2 Quantity of Investments

5.2.1 Geographic Expansion

I now turn to the intensive margin of venture capital investment and examine whether data-driven VCs invest in more startups in non hub locations than they did before technology adoption. I construct a VC-year panel, still focusing on hub based investors, and examine the number of investments they make in non-hub locations each year. More formally, I estimate the following regression at the VC-year level:

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

The dependent variable is the number of non-hub investments made by VC j in year t. The main explanatory variable, $DataDriven$, is an indicator equal to one if VC j is classified as data-driven in year t and zero otherwise. The vector $X_{j,t}$ includes time-varying controls for VC j, such as firm age, number of employees, and total assets under management. To account for

time-invariant firm characteristics, I include VC fixed effects, α_j. In addition, I incorporate state $\times$ industry focus $\times$ funding stage $\times$ year fixed effects, $\alpha_{c\times i\times s\times t}$, to mitigate concerns that geographic, industry-specific, or temporal shocks coinciding with data technology adoption may drive investment patterns outside major hubs. The coefficient of interest, β, is therefore identified by comparing changes in VC j's non-hub investments before and after adoption relative to other VCs operating within the same state-industry-stage-year segment. Given that the outcome variable is a count measure that is left-censored at zero and exhibits positive skewness, I estimate the model using a Poisson specification, clustering standard errors at the VC-firm level.

Table 5 reports the results. Column (1) estimates Equation 4 with investor age as the sole control. The coefficient on $DataDriven$ is positive and statistically significant, indicating that, following adoption of data technologies, VCs increase the number of investments made in non-hub regions by approximately $e^{0.336} - 1 \approx 40\%$ relative to other VCs in the same state-industry-stage-year segment. Column (2) adds size controls including the natural logarithm of assets under management and the number of employees. The coefficient on $DataDriven$ remains statistically significant at the 5% level, though its magnitude declines to 0.183, implying that VCs increase their investments in non-hub locations by roughly 20%. Given that the average VC in the sample makes two investments per year in non-hub regions, this effect corresponds to an additional 0.4 investments annually. The reduction in magnitude after including size controls is consistent with the view that venture capital is difficult to scale: individual partners can monitor only a limited number of portfolio companies, so larger firms with more resources tend to make more investments (Gompers and Lerner (2001)).

One concern is that VCs may be more likely to hire data scientists during periods of expansion, when they are already diversifying the geographic scope of their investments. To address this issue, I examine pre-trends. Specifically, I implement a stacked difference-in-differences (DiD) design. Because the timing of data-scientist hiring is staggered across VCs, 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 five years before and after the hiring event. The panel is unbalanced because some VCs were founded fewer than four years prior to the hiring event. The analysis is conducted at the VC-year level. I estimate the dynamics by interacting event-time dummies with treatment status using the following specification:

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

where $Treated_{dc}$ is an indicator equal to one if investor j in cohort d eventually becomes data-driven, and $Year(k)_c$ indexes years relative to cohort d's hiring of a data scientist. Panel A of Figure 2 presents the results for the number of non-hub investments made by VC j in year t. There are no significant pre-trends, suggesting that data technology adoption is not preceded by an upward trend in non-hub investment activity.

Another concern is that, while this regression design shows that the number of investments in non hub areas increase post technology adoption, it does not show how much their non-hub investments increase compared to total investments. To address this concern, I create a stacked sample at the VC-hub-year level:

$$Y_{j,h,t} = \beta_1 DataDriven_{j,t} + \beta_2 DataDriven_{j,t} \times NonHub_{h,t} + \beta_3 NonHub_{h,t} + X_{j,t} + \alpha_j + \alpha_{c\times i\times s\times t} + \epsilon_{j,t} \quad (4)$$

The results are in column (3) of Table 5. The coefficient on $DataDriven$ is 0.135, indicating that after adopting data technologies, VCs increase the number of investments in hub regions by approximately $e^{0.135} - 1 \approx 14\%$, relative to non-data-driven peers in the same state-industry-stage-year segment. The interaction term, $DataDriven \times NonHub$, is 0.137 and statistically significant, suggesting that adoption leads to a disproportionately larger increase in non-hub investments. Interpreted together, the estimates imply that data-driven VCs expand their hub activity by roughly 14% while increasing their non-hub investments by about $e^{0.135+0.137} - 1 \approx 30\%$. In column (4), I include VC-firm-by-year fixed effects, which absorb $DataDriven$, $NonHub$, and all time-varying VC controls. The coefficient on $DataDriven \times NonHub$ remains stable at 0.137, indicating that the additional increase in non-hub investments relative to hub investments is not driven by time-varying VC characteristics, but rather reflects a systematic tilt in portfolio geography following data technology adoption. Overall, these results show that while VCs modestly expand their investments overall after adoption, data-driven VCs also shift the composition of their portfolios toward startups located outside of major hubs.

5.2.2 Industry Expansion

Beyond geography, VCs also tend to specialize in particular industries (Hochberg, Mazzeo, and McDevitt (2015)), which can limit exposure to novel opportunities. Because industry specialization is less binary than office location, I use the VC-year panel to examine whether data-driven investors expand into industries outside of their primary focus after adopting data technologies. VCs are classified into 12 main industry groups[20] and, on average, invest in startups in five different industries per year. Thus, while most VCs maintain a primary industry focus, there is often considerable overlap across related industries. For instance, a VC whose main classification is software may also invest frequently in internet services and artificial intelligence. I estimate the same specification as Equation 4, using the full sample of VCs. The results are reported in Table 6.

Columns (1) and (2) consider the number of industries a VC invests in during a given year outside of its main industry. In column (1), without VC-firm fixed effects, the comparison is between data-driven and non-data-driven VCs operating in the same state-industry-stage-year segment. The coefficient on $DataDriven$ is positive and statistically significant, implying that data-driven VCs invest in approximately $e^{0.21} - 1 \approx 23\%$ more industries than their non-data-driven peers. Given that the average VC invests in 4.3 industries outside its primary focus each year, this corresponds to roughly one additional industry. Column (2), which includes VC-firm fixed effects, shows that the effect remains positive and significant, indicating that even within VCs, the number of industries invested in per year increases following adoption of data technologies.

In columns (3) and (4), the dependent variable is the number of investments a VC makes in startups outside of its primary industry. The coefficients are positive and statistically significant across VCs (column (3) and within data-driven VCs (column (4)). The estimate in column (4) indicates that, after adopting data technologies, VCs increase their investments in startups outside their main industry by approximately 13 percent.

Finally, columns (5) through (8) focus on VC investments in industries that are typically further removed from the main technology focus of most data-driven firms, namely health (health care and biotechnology) and energy and manufacturing and sustainability. If data technologies lower search frictions, the largest increases should occur in industries outside a VC's established

20. The 12 industries are: professional services, media, commerce, biotechnology, health care, software, internet services, artificial intelligence, financial services, energy and sustainability, semiconductors, and cybersecurity.

expertise. The within-VC estimates confirm this prediction: after adoption, VCs increase their number of investments in health care by 37 percent (column (6)) and in manufacturing and sustainability by 19 percent (column (8)). Taken together, these results indicate that the adoption of data technologies broadens the scope of VC investment activity, enabling investors not only to diversify beyond their primary industry focus but also to expand into sectors that are traditionally more distant from their established expertise.

5.3 Robustness

This section shows that the main results are robust to a variety of checks: restricting the analysis to first-time financings; using alternative geography definitions; reestimating the analyses to account for the Covid-19 pandemic; and employing alternative constructions of $DataDriven$.

5.3.1 First Time Investments

One concern with the main analysis is that it includes all investments, not just initial financings. Data technologies are most likely to matter at the discovery stage, when a VC invests in a startup for the first time. I therefore reestimate Equation 1 using only first-time investments by a VC in a given startup, with counterfactual matches restricted to first-time financings as well. Table A1 reports the results. Column (1) includes the baseline specification with all controls and yields a larger coefficient, corresponding to a 6% increase in the likelihood of a data-driven VC investing in a non-hub startup. Columns (2) through (4) show heterogeneity across founder types, with stronger effects for founders located in low-activity commuting zones, educated outside the top 20 universities, and first-time founders. Together, the results confirm that the main findings are not driven by follow-on investments.

I conduct a similar analysis for the intensive margin tests, reestimating Equation 4 using the number of first-time investments by a VC in a given startup as the dependent variable. Table A2 reports the results: column (1) covers first-time investments in non-hub startups, column (2) covers first-time investments outside the VC's main industry focus, column (3) focuses on health care and biotechnology, and column (4) focuses on energy and sustainability. The coefficients are quantitatively similar across specifications, and in columns (1) through (3) they are statistically

significant at the 1 or 5 percent level, confirming that the post-adoption expansion is robust at the initial investment margin.

5.3.2 Different Geography Definitions

In Section 5.2, I show that after VCs adopt data technologies, they increase the number of investments they make in non-hub startups. A concern is that this increase could be concentrated in a single non-hub location, which would not necessarily indicate reduced search frictions if it reflects a preexisting or growing local network. To address this, I replace the dependent variable in Equation 4 with the number of distinct non-hub commuting zones a VC invests in within a given year. Column (1) of Table A3 reports a positive and statistically significant coefficient on $DataDriven$, indicating that VCs expand the number of non-hub locations they invest in after adopting data technologies.

A second concern is that VCs may be shifting toward non-hub areas that already have active VC markets. Although approximately two thirds of VC activity occurs in San Jose and San Francisco, California, Boston, Massachusetts, and New York, New York, several other regions have a growing VC presence. I define eleven additional hubs ("secondary hubs") to capture these markets. If post-adoption investments are concentrated in secondary hubs, the pattern could be driven by professional networks that data-driven VCs have with local VCs rather than technology. In columns (2) through (4) of Table A3, I split non-hub investments into secondary hubs (column 2), non-secondary hubs (column 3), and low-activity commuting zones (column 4), where low activity is defined as commuting zones that received 25 or fewer investments over the prior five years. Across all specifications, the coefficients are positive and statistically significant, with the largest effect in the low-activity commuting zone specification, indicating the greatest post-adoption increase occurs in places with the least VC presence. This mitigates the concern that the non-hub expansion is limited to secondary hubs.

5.3.3 Covid-19

Another concern is omitted-variable bias from the Covid-19 pandemic. Prior work shows that, after 2020, VCs increased investment in geographically distant startups because limits on face-to-face interaction constrained the in-person monitoring that is central to venture investing (Han

et al. (2025), Dalla Fontana et al. (2025)). Over roughly the same period, rapid advances in AI made data tools more accessible, including the broad release of ChatGPT in 2022. As a result, data technology adoption and a shift toward out-of-network investing may be correlated with the pandemic rather than technology driving the change. To address this concern, in Table A4 I reestimate the main specifications restricting the sample to end in 2019. The coefficients are slightly smaller in magnitude, but the results remain robust.

5.3.4 Different Measures of Data-Driven

The main independent variable is a binary indicator that equals one beginning in the year a VC hires its first data-related employee. However, this measure ignores the intensity and subsequent growth of the VC's data function. The estimated increase in data-driven investments in non-hub locations and out-of-industry deals remains robust when I replace the indicator with the natural log of one plus the number of data employees (Panel A) and when I scale the number of data employees by the number of partners (Panel B) in Table A5. The log specification addresses skewness and captures diminishing marginal effects of additional hires, whereas the partner-scaled measure normalizes data capacity for organizational size and decision-making bandwidth, improving comparability across firms.

5.4 Selection: Instrumental Variables Approach

The previous section presented evidence that after VCs adopt data technologies, they tend to invest more in startups outside of their professional networks, as proxied by geography and industry specialization. A concern, however, is that unobservables such as pipeline quality, strategic shifts, or changes in geographic focus could simultaneously influence both adoption and non-hub investing. Standard instruments for technology adoption used in other settings are either limited or invalid in the VC context. For example, the highly concentrated nature of VC activity makes it difficult to exploit geographic variation, such as broadband penetration (Akerman, Gaarder, and Mogstad (2015)), while the concentration of VC activity within a few universities leaves little usable variation in alumni hiring networks (Babina et al. (2024)). Even within the VC literature, using the AWS infrastructure shock (Bonelli (2025)) as an instrument is problematic in

my setting.[21]

My identification strategy therefore focuses on the *timing* of technology adoption, interacted with a predetermined, VC-specific, AI exposure measure. This shift-exposure structure isolates plausibly exogenous variation in adoption timing and mitigates concerns that contemporaneous demand shocks drive both technology adoption and investment strategies.

5.4.1 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 idea is that commercial interest in AI only became widespread around 2010,[22] when technology firms first began incorporating AI into consumer products (e.g., Apple introducing Siri in 2011) and non-technology firms started applying AI in business operations (e.g., Walmart using AI-enabled cameras in 2017). VCs with greater exposure to AI prior to this period are therefore more likely to adopt data technologies earlier. The second source of variation is the timing of raising a new fund, since VCs often expand personnel at fundraising events and data scientists represent high-cost, specialized hires (Gompers and Lerner (2001)). The interaction of these two components generates quasi-exogenous variation in the timing of data technology adoption.

AI Industry Exposure Recent research by Babina et al. (2024) documents a sharp rise in AI adoption by public firms across industries, driving growth in sales, employment, and valuations. For non-technology sectors such as financial services, successful adoption required both awareness of AI's benefits and the expertise to implement it. While AI only gained widespread commercial use after 2010, several startups were pioneering applications in the 2000s. For example, Predictix (founded in 2005) used big data analytics to forecast business operations, and Voci (founded in 2008) developed speech-to-text algorithms in hardware. VCs backing such startups gained early exposure to AIs potential applications, giving them a first-mover advantage. I hypothesize that VCs with AI-focused investments prior to 2010 are more likely to adopt data

21. The intuition is that the introduction of AWS lowered the cost for certain startups to enter, which in turn increased the incentives for VCs in those industries to adopt data technologies to better manage the pre-investment screening process. However, the AWS shock may also have disproportionately reduced entry costs for startups in non-hub regions or specific industries, directly correlating with both VC adoption and my main outcome variables.

22. Medium : How AI Evolved: A Timeline of Artificial Intelligence from 1950 to 2025.

technologies earlier, as they would already possess the knowledge and experience to recognize their value. Importantly, the exclusion restriction is likely to hold in this setting, as investments in AI startups prior to 2010 would not have been motivated by an intent to adopt data technologies themselves, given that widespread commercial adoption had not yet begun.

I exploit cross-sectional heterogeneity in exposure to AI-related industries to identify the effect of AI experience on VC adoption of data technologies. Crunchbase classifies startups into over 750 market categories, allowing me to account for fine-grained differences in startup focus.[23] Using AI-related terms defined in the Artificial Intelligence Glossary from Tech Target, a marketing company that provides data-driven services to B2B technology vendors,[24] I identify which of the 750 Crunchbase categories are AI-related. Table B1 reports the resulting set of categories. For each VC firm founded before 2010, I compute the share of its investments between 2000 and 2009 that were in AI-related categories. Because VC firms' capacity to invest is constrained by organizational scale (Gompers and Lerner (2001)), I scale this measure to avoid automatically assigning greater intensity to larger firms that make more investments overall. Specifically, I assign each VC an exposure score based on its percentile rank in the distribution of AI investment shares:

$$VCExposure_j = \text{Rank}_I \left(\frac{\text{AI Investments}j}{\text{Total Investments}_j} \right). \tag{5}$$

Fund Timing Next, I examine the timing of new fundraisings to identify periods when VCs are most likely to hire new employees. In my sample, VCs raise a new fund approximately every 4.5 years, and I focus on funds raised from 2010 onwards. As a first step, using the VC-year panel regression (Equation 4), I document that firms tend to increase hiring between the year of raising a new fund and two years after (Table B2).

A potential concern is that VCs choose to raise a new fund in response to potential opportunities in out-of-network areas, such as non-hub geographies or industries outside their historical specialization, which would violate the exclusion restriction. However, the timing of new fundraisings is not fully endogenous to the VC firm, as GPs cannot freely decide when to raise. Fundraising is determined by a combination of factors such as the performance of the prior fund

23. For example, within the broader segment *Financial Services*, Crunchbase includes Life Insurance, FinTech, Mobile Payments, and Wealth Management.

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

(Lerner, Schoar, and Wongsunwai (2007); Harris et al. (2023)), demand for venture allocations by limited partners (Gompers and Lerner (1999)), and broader public market conditions (Kaplan and Schoar (2005)). Funds are structured with finite lives, and LPs typically recommit on contractual schedules, making fundraising cycles more rigid than opportunistic. An additional concern, however, is that fundraising cycles might still coincide with geography- or industry-specific trends, which could correlate out-of-network investment opportunities with local or sectoral shocks. To address this, I calculate the correlations in fundraising cycles among two highly concentrated groups: 10 randomly selected San Jose-based software VCs and 10 randomly selected Boston-based biotechnology VCs. As reported in Table B3, the timing of fundraising within these industry-geography clusters is largely uncorrelated, with average correlations of -0.08 for both groups.

These tests, together with the institutional features of fundraising, suggest that the timing of raising a new fund is (1) positively associated with new employee hires and (2) not systematically correlated with local VC trends, making it a useful source of variation for instrumenting data scientist hiring. Based on this, I construct an indicator equal to one if a VC raised a fund in the previous two years ($FundRaised_{j,[-2:0]}$). Data on VC fund vintages are collected from Preqin and Crunchbase.

Instrument Since the instrument focuses on the timing of data technology adoption, I construct an indicator that captures when VCs are most likely to adopt data technologies. The instrument leverages differential exposure to AI prior to 2010, interacted with the arrival of a new fund, conditional on the VC not yet having adopted:

$$Z_{j,t} = \left(VCExposure^{\text{pre-2010}} j\right) \times \mathbf{1}\{FundRaised_{j,t-2:t}\} \times \mathbf{1}\{D_{j,t-1} = 0\}. \tag{6}$$

where $D_{j,t}$ is a persistent indicator equal to one after the firm first adopts data technologies and remaining one thereafter (once adopted, never reverted)[25]. Intuitively, $Z_{j,t}$ generates quasi-exogenous variation in the timing of adoption. VCs with higher pre-2010 exposure to AI are more likely to understand and adopt related technologies, and the fundraising cycle provides a natural

25. In the data, no VC ever reverts to being "non-data-driven." Moreover, most VCs employ only one to three data scientists, so there is no meaningful "intensity" margin that varies smoothly year to year, supporting the absorbing-treatment interpretation.

period for organizational retooling. The instrument is gated on firms that have not yet adopted, ensuring it only shifts the hazard of adoption for those still untreated.

The instrument shares similarities with a shift-share design (Bartik (2002), Adao, Kolesár, and Morales (2019), Borusyak and Hull (2023), Goldsmith-Pinkham, Sorkin, and Swift (2020)). In this context, the "share" is given by a VC's pre-2010 exposure to AI startups, which captures heterogeneity in how likely each firm is to benefit from or understand the technology prior to its widespread commercialization. The "shift" is the arrival of a fundraising event, which provides a quasi-exogenous opportunity for organizational restructuring and the hiring of specialized personnel, including data scientists. The interaction of these two components generates variation in the timing of adoption across VCs, conditional on not yet having adopted. The identifying assumption is that, conditional on fixed effects and controls, the timing of fundraising is not systematically related to unobserved investment opportunities, and the heterogeneity in pre-2010 AI exposure captures only differential propensity to adopt rather than differential investment trends. This assumption ensures that the instrument isolates plausibly exogenous shifts in adoption timing, rather than reflecting underlying demand shocks for non-hub investments.[26]

5.4.2 The Causal Impact of Data Technology Adoption and VC Investments

First Stage I begin by estimating the causal impact of data technology adoption on VC investments in non-hub locations, focusing on intensive margin tests. I therefore restrict my sample to VCs located in hub locations. Because the instrument is designed to capture the *timing* of adoption, I construct an at-risk indicator that flags periods when VCs are likely to adopt, rather than relying on the persistent $DataDriven$ measure. This indicator reflects the hazard of adoption, defined as:

$$Adopt_{j,t} = \mathbf{1}\{D_{j,t} = 1 \;\&\; D_{j,t-1} = 0\}, \tag{7}$$

where $D_{j,t}$ is a persistent indicator equal to one after a VC first adopts data technologies and remaining one thereafter (once adopted, never reverted).

26. A common concern in shift-share designs is that the "shares" may themselves be endogenous, potentially biasing the instrument (Goldsmith-Pinkham, Sorkin, and Swift (2020)). In this setting, the use of pre-2010 AI exposure mitigates such concerns, as it is predetermined relative to the post-2010 surge in AI commercialization and adoption.

I instrument the adoption hazard of VC j with its exposure to AI prior to 2010 interacted with whether the VC raised a new fund in the previous two years, conditional on not having adopted beforehand. The first-stage specification is:

$$Adopt_{j,t} = Z_{j,t} + \alpha_j + \alpha_{c\times i\times s\times t} + \epsilon_{j,t}, \tag{8}$$

where $Z_{j,t}$ captures heterogeneity in adoption hazards by shifting the probability of adoption when a VC is most likely to reorganize (around fundraising events), but only among firms that have not yet adopted. α_j are VC-firm fixed effects, $\alpha_{c\times i\times s\times t}$ are year×state×stage×industry effects, and $X_{j,t}$ (not shown) includes time-varying VC controls.

Column (1) of Table 7 reports the first-stage results. The coefficient on $Z_{j,t}$ is positive and statistically significant, and the F-statistic of 22.23 exceeds the conventional threshold of 10, satisfying the instrument relevance condition. To assess observable pre-trends, I also estimate the following event study specification:

$$Adopt_{j,t} = \sum_{k=-3,k\neq-1}^{3} \beta_k\{VCExposure_j^{\text{pre-2010}} \times Year(k)_t\} + X_{j,d,t} + \alpha_j + \alpha_{c\times i\times s\times t} + \epsilon_{j,t} \tag{9}$$

where $Year(k)_t$ is a dummy variable equal to one if year t corresponds to k years before or after a fundraising event. I focus on a window of three years before and three years after, as most VCs raise new funds every three to five years, making this the cleanest horizon for analysis. Panel A of Figure 3 plots the estimated coefficients on β_k. The results show no evidence of pre-trends. The likelihood of a VC adopting data technologies rises immediately following the fundraising event, remains elevated for the subsequent two years, and then converges back to zero by year $t+3$. This pattern indicates that hiring is closely tied to the fundraising event and concentrated in its immediate aftermath.

Reduced Form Since adoption is measured as a short-run event and only a small share of firms adopt during the sample (fewer than 10%), the IV ratio based on a narrow complier group can be noisy and inflate coefficient magnitudes (Angrist, Imbens, and Rubin (1996), Angrist and Pischke (2009), Stock and Yogo (2002)). I therefore prioritize the reduced form (intention-to-treat),

which directly links the instrument to investment outcomes without dividing by a potentially weak or transitory first stage:

$$Y_{j,t} = \beta Z_{j,t} + X_{j,t} + \alpha_j + \alpha_{c \times i \times s \times t} \tag{10}$$

The results are displayed in column (2) of Table 7. The coefficient on $Z_{j,t}$ is positive and statistically significant, indicating that early exposure to AI prior to 2010, when interacted with a fundraising window, predicts subsequent non-hub investments. The estimated magnitude is 0.261. Interpreted as an *intention-to-treat* effect, this implies that moving from the 25th to the 75th percentile of pre-2010 AI exposure during a fundraising window increases the expected number of non-hub investments by approximately $e^{0.261 \times 0.50} - 1 \approx 13\%$. Put differently, for VCs with greater pre-2010 AI exposure, the arrival of a new fund corresponds to a disproportionate reallocation of activity toward non-hub regions. Columns (3) and (4) present results from a stacked regression at the VC-hub level. These estimates are quantitatively similar to those reported in columns (3) and (4) of Table 5, suggesting that VCs with higher AI exposure increase their non-hub investments around fundraising events relative to their hub investments. Panel B of Figure 3 shows no noticeable pre-trends, reinforcing the validity of the reduced-form specification. In addition, I lag the number of non-hub investments by one to four years in an additional specification reported in Table B5. Across all horizons, the coefficients remain positive and statistically significant, indicating that adoption has lasting effects on VCs' propensity to increase non-hub investments beyond the immediate fundraising window.

Discussion of Instrument While I report the second-stage IV estimates in the Table B4 (which are positive and statistically significant) for completeness, I emphasize the reduced-form results in the main text. Because adoption is rare (fewer than 10% of firms adopt during the sample), the IV ratio is identified off a narrow complier group, which can inflate the estimated coefficients and produce effects that are difficult to generalize (Bound, Jaeger, and Baker (1995), Stock, Wright, and Yogo (2002)). In particular, the local average treatment effect (LATE) applies only to those VCs whose adoption is shifted by the instrument, rather than to the broader population of always- or never-adopters, and thus tends to magnify the magnitude of the causal effect (see Appendix B.1 for a fuller discussion; Angrist, Imbens, and Rubin (1996)). By contrast, the re-

duced form captures the intention-to-treat effect, linking quasi-exogenous shifts in exposure and fundraising directly to investment outcomes. This approach avoids the scaling issues inherent in the IV ratio (Nelson and Startz (1988)).

Impact on Investing Outside Industry I also estimate the reduced-form specification (Equation 10) to test whether adopting data technologies causally expands the industry scope of VC investments. In this analysis, I use the full sample of VCs. The results are displayed in Table 8. The first stage remains positive and statistically significant with an F-statistic of 19.4 (column (1)), supporting the relevance condition.

Columns (2) and (3) examine the number of industries outside a VC's main focus and the total number of investments in those industries, respectively. In both cases, the coefficient on $Z_{j,t}$ is positive and statistically significant, with magnitudes quantitatively similar to the OLS results reported in Table 6. These results suggest that VCs with greater pre-2010 AI exposure increase their investments outside their focal industry after raising a new fund.

In columns (4) and (5), I focus on investments in the health care industry and the energy and sustainability industry, respectively. Since data-driven VCs are concentrated in the technology sector, I would expect these two industries to experience the largest increases if data technologies reduce information frictions. Consistent with this prediction, the coefficients are positive and statistically significant, indicating a causal effect.

For completeness, I also report the second-stage regressions, where $Adopt$ is instrumented with $Z_{j,t}$. The results, displayed in Table B6, are directionally consistent but substantially larger in magnitude. This reflects the fact that the IV ratio is identified off a narrow complier group, yielding a local average treatment effect that is not directly generalizable.

6 Data-Driven Performance

In the previous section, I find that data-driven VCs are more likely to invest in out-of-network startups, as proxied by non-hub and out-of-industry ventures. In this section, I examine the quality of such data-driven investments. Prior literature shows that adopting data technologies provides an advantage in identifying startups that are less likely to fail (i.e., more likely to receive follow-on funding) but does not improve the ability to find firms that will achieve a major exit,

either through an IPO or an acquisition (Bonelli (2025)). However, little is known about how data-driven investments in out-of-network startups compare to (1) their data-driven counterparts or (2) traditional investments in out-of-network deals.

Ex ante, the performance of data-driven investments outside a VCs' network is ambiguous. On the one hand, VCs may be less effective monitors for startups outside of their focus. Monitoring is an important part of the VC investment process, and is associated with stronger performance (Bernstein, Giroud, and Townsend (2016)). For example, if non-hub investments are geographically distant, VCs may be less able to effectively manage startups of similar quality as their hub investments. On the other hand, investing in areas with less competition like non-hubs may provide access to higher-quality opportunities or more favorable terms (Hochberg, Mazzeo, and McDevitt (2015)). In addition, VCs may also apply stricter selection criteria to deals farther from their focus, knowing that active monitoring will be more difficult (Chen et al. (2010)). Thus, the performance of data-driven investments outside a VC's main focus remains an empirical question.

I therefore examine the performance of data-driven investments compared to traditional VCs. I split performance outcomes into two main categories: if the startup receives another round of financing ($Follow\ On$) and if the startup achieves a major success outcome either through a high quality acquistion or IPO ($Major\ Exit$). I examine this at the investment level and conduct the following regression:

$$Y_{k,j,t} = \beta\, DataDriven_{j,t} \times Out\text{-}of\text{-}Network_{k,t} + X_{k,j,t} + \alpha_j + \alpha_{c\times i\times s\times t} + \epsilon_{i,j,t}$$

The dependent variable is an indicator if the startup receives follow on ($FollowOn$) funding or eventually exits through an IPO or acquisition ($Major\ Exit$). $DataDriven_{j,t}$ is an indicator equal to 1 from the date a VC hires their first data-related employee. $Out\text{-}of\text{-}Network_{k,t}$ is an indicator if the startup is located in a non hub commuting zone ($Non\ Hub$) or operates in a different industry than the VCs' main focus ($Diff\ Industry$). $X_{k,j,t}$ includes various VC, startup and VC-startup time varying controls. α_j are VC firm fixed effects to absorb for any time invariant VC characteristics and $\alpha_{c\times i\times s\times t}$ compares VCs investing in startups inm the same industry-state-stage-year segment.

6.1 Non-Hub Investments and Follow-On Funding

I begin by examining whether data-driven investments are more likely to receive follow-on funding. The sample is restricted to investments made between 2010 and 2022, since follow-on rounds typically occur within two years. The results are displayed in Table 9. Consistent with prior evidence, column (1) shows that data-driven investments are more likely to receive follow-on funding (Bonelli (2025)). The coefficient on $DataDriven$ implies that, after VCs adopt data technologies, the probability of their portfolio companies receiving additional funding rises by 1.3 percentage points, or about 2.2% relative to the unconditional mean.

In column (2), I restrict my sample to investments made by VCs located in hub commuting zones. I interact $DataDriven$ with an indicator equal to 1 if the startup is located in a non-hub region ($NonHub$). The coefficient on $DataDriven \times NonHub$ is negative and statistically significant at the 10% level, while the coefficient on $DataDriven$ remains positive. The difference in the coefficients are significant at the 10% level. Thus data-driven investments in non-hubs are less likely to receive follow-on financing than their hub counterparts. However, when compared to the coefficient on $NonHub$, which captures traditional investments in non-hubs, the difference is not significant. This indicates that data-driven investments in non-hubs behave similarly to traditional non-hub investments. The coefficient on $DataDriven \times NonHub$ remains consistent and increases in statistical significance when including VC-by-year fixed effects, which controls for all unobserved heterogeneity within the VC firm in a given year.

In column (4), I focus on VC investments in different industries. I use the full VC sample. The coefficient on the interaction term ($DataDriven \times DiffIndustry$) is largely insignificant and is not statistically different from data driven investments in the same industry ($DataDriven$) and traditional investments outside the VCs focal industry ($DiffIndustry$). The coefficient on $DataDriven \times DiffIndustry$ remains the same when including VC-by-year fixed effects (column (5)).

Overall, the results suggest that while data technologies improve VCs' ability to identify startups that receive follow-on funding, this advantage does not extend to out-of-network deals. In particular, data-driven investments in non-hub regions are less likely to receive follow-on financing than comparable hub deals, consistent with the idea that geographic distance makes monitor-

ing more difficult. Moreover, these non-hub data-driven deals perform no better than traditional non-hub investments, indicating that data technologies do not offset the monitoring challenges associated with investing outside of a VC's main focus.

6.2 Non-Hub Investments and Major Exit

The previous analysis shows that data-driven investments in out-of-network startups are less likely to receive follow-on funding, consistent with the idea that monitoring such deals is more difficult and that VCs are more likely to abandon them at earlier stages. However, VCs ultimately generate most of their returns through major acquisitions or IPOs. While out-of-network investments may be harder to monitor, data technologies could still provide an advantage at the selection stage by lowering search frictions and identifying promising startups in regions with less competition from other investors. Major exits are also rare (27% in my sample), so they capture the most consequential outcomes for VC performance. To examine this, I restrict the sample to startups that survive to at least a Series B round, which filters out early failures and allows me to focus on those with stronger growth potential. If VCs are more likely to abandon out-of-network investments because of higher monitoring costs, then these investments may fail to reach later rounds even if their initial quality is comparable to in-network startups. By conditioning on survival to Series B, I focus on startups that have already demonstrated viability and secured multiple rounds of external financing. This ensures that the comparison between hub and non-hub investments is made among startups with similar prospects, rather than confounding monitoring challenges with differences in early-stage survival.

The results are reported in Table 10. In column (1), the coefficient on $DataDriven$ is negative and statistically significant, indicating that data-driven investments are 2.9 percentage points less likely to achieve a major exit, an 11% decrease from the unconditional mean. Taken together with the positive effect on follow-on funding reported in column (1) of Table 9, this pattern is consistent with prior evidence that data technologies are more effective at screening startups out failures, but less effective at identifying novel innovations that generate major outcomes (Bonelli (2025)).

In column (2), I interact $DataDriven$ with $NonHub$. The coefficient on $DataDriven \times NonHub$ is positive and statistically significant, and it is statistically different from the coefficient

on $DataDriven$. This indicates that data-driven investments in non-hub startups are more likely to exit through IPO or acquisition than comparable hub investments. Moreover, the coefficient on the interaction is statistically different from the coefficient on $NonHub$, showing that data-driven non-hub investments are more likely to succeed than traditional non-hub investments. In column (3), I include VC-by-year fixed effects. The coefficient on $DataDriven \times NonHub$ remains positive and similar in magnitude, although it becomes statistically insignificant.

In columns (4) and (5), I replace $NonHub$ with $DiffIndustry$. The coefficient on $DataDriven \times DiffIndustry$ is positive and statistically significant, indicating that data-driven investments outside a VC's main industry focus are more likely to achieve major exits than investments within their core industry. In addition, the coefficient on $DiffIndustry$ is negative and statistically significant at the 1% level, showing that data-driven investments outside a VC's main industry outperform traditional out-of-industry investments.

Taken together with the follow-on funding results, these findings suggest that while data technologies broaden the scope of investment opportunities for VCs, their main advantage lies in identifying promising startups outside of their specializations. Conditional on surviving the early stages, these out-of-network data-driven investments are the most likely to achieve major exits. Furthermore, they outperform traditional out-of-network investments, underscoring the role of data technologies in lowering search frictions and uncovering high-potential ventures.

7 Implications for Areas with Low History of VC Activity

To summarize the results so far, I show that after adopting data technologies, VCs are more likely to invest outside of their established networks, as proxied by non-hub and out-of-industry investments, and that these investments are disproportionately likely to succeed through IPO or unicorn outcomes. Up to this point, the analysis has mainly focused on outcomes for the VCs that adopt data technologies. A natural follow-up question is whether certain regions or sectors could also benefit from this more systematic approach to deal sourcing. VC activity is heavily concentrated in a few major hubs, meaning that areas with historically low VC presence may stand to gain the most. A data-driven VC entering such a market can increase the visibility of local startups, both by directly expanding the set of firms considered and by indirectly placing

the region within the radar of other investors, creating potential spillover effects. In this section, I examine whether areas that receive data-driven investment subsequently experience broader increases in VC activity. I begin by outlining the data and research design before presenting the results.

7.1 Data and Research Design

I begin by constructing a panel of commuting zones that have historically received little VC activity ("low-activity" hubs). I define low activity hubs as a commuting zone with 25 or fewer VC investments over the previous five years (Hochberg, Ljungqvist, and Lu (2010)). Treated zones are those where a data-driven VC makes its first local investment, totaling 56 commuting zones. All other low activity zones serve as the control group.

To account for local entrepreneurial dynamics, I include commuting zone-level controls for entrepreneurial activity and innovation. In particular, I use data from the "Startup Cartography Project" (SCP; Fazio et al. (2019)), which provides entrepreneurial ecosystem statistics for the United States from 1988 to 2016 at multiple geographic levels. The SCP combines state-level business registration records with predictive analytics to estimate the probability of "extreme" growth (IPO or high-value acquisition) at or near founding for all new firms in a given year. These estimates are used to construct four ecosystem statistics: the rate of startup formation, average entrepreneurial quality, quality-adjusted startup activity, and ecosystem performance over time. My primary measure is the *Regional Entrepreneurship Cohort Potential Index* ($RECPI$), which interacts the number of new business registrants with the expected growth potential of those startups. In short, the $RECPI$ is a quality-adjusted measure of entrepreneurial activity in a given area. In addition, I include income, GDP, and the share of the population with a college degree to capture broader economic conditions.

I then estimate a stacked difference-in-differences model, comparing changes in VC activity before and after data-driven VCs enter a commuting zone. The baseline specification is:

$$Y_{d,c,t} = \beta Treated_{d,c} \times Post_{d,t} + \alpha_{d,c} + \alpha_{d,t} + \epsilon_{d,c,t}, \tag{11}$$

where $Y_{d,c,t}$ denotes outcomes of VC activity for commuting zone c in cohort d and year t.

$Treated_{d,c}$ equals one if a startup in zone c receives an investment from a data-driven VC, and $Post_{d,t}$ equals one after entry and zero otherwise. The specification includes cohort-commuting zone fixed effects ($\alpha_{d,c}$) to absorb time-invariant differences across zones and cohort-year fixed effects ($\alpha_{d,t}$) to capture common time shocks. I further interact $Post_{d,t}$ with baseline commuting zone characteristics (income, GDP, and college share) to allow for heterogeneous post-entry dynamics across regions, and in the tightest specification, I also interact $Post_{d,t}$ with $\log(RECPI)$ to account for pre-existing entrepreneurial activity.

Because the outcomes are nonnegative, left-censored at zero, and highly skewed, I estimate the model using Poisson regressions. The coefficient of interest, β, measures the change in VC activity in commuting zones receiving a data-driven investment relative to those that do not.

7.2 Results

I examine the impact of an investment by a data-driven VC in a commuting zone with low VC activity on entrepreneurial outcomes in subsequent years. Specifically, I study the number of funding rounds, the number of startups receiving their first VC financing, the number of unique investors, the number of first-time investors, and the number of patents filed by VC-backed startups. To classify patents, I follow the methodology introduced by Ewens and Marx (2024), which identifies patents associated with VC-backed firms. I estimate the specification in Equation 11, and results are reported in Table 11.

Columns (1) and (2) show that the number of funding rounds increases by 8-12% in commuting zones that receive a data-driven investment relative to those that do not. Columns (3) and (4) similarly show an 8-12% increase in the number of startups receiving their first VC financing. In columns (5) and (6), the number of unique investors rises by 19-26%, while columns (7) and (8) indicate that the number of first-time investors increases by 11-18%. Finally, columns (9) and (10) show that patents filed by VC-backed startups increase by 28-35% following entry by a data-driven VC.

Taken together, these results highlight that data-driven entry is associated with meaningful increases in entrepreneurial activity, investor participation, and innovation in commuting zones with historically limited VC presence. This suggests that data technologies not only reshape VC strategies but also broaden the geography of entrepreneurial finance, potentially aiding founders

who would otherwise have limited exposure to VC networks.

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. 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 and their main industry focus. 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.

Adao, Rodrigo, Michal Kolesár, and Eduardo Morales. 2019. "Shift-share designs: Theory and inference." *The Quarterly Journal of Economics* 134 (4): 1949–2010.

Akerlof, George A. 1970. "The market for âlemonsâ: Quality uncertainty and the market mechanism." In *Uncertainty in economics,* 235–251. Elsevier.

Akerman, Anders, Ingvil Gaarder, and Magne Mogstad. 2015. "The skill complementarity of broadband internet." *The Quarterly Journal of Economics* 130 (4): 1781–1824.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91 (434): 444–455.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

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.

Bartik, Timothy J. 2002. *Instrumental variable estimates of the labor market spillover effects of welfare reform.* Technical report. Upjohn Institute working paper.

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. 2025. "Data-driven Investors." In *Data-driven Investors: Bonelli, Maxime.* [Sl]: SSRN.

Borusyak, Kirill, and Peter Hull. 2023. "Nonrandom exposure to exogenous shocks." *Econometrica* 91 (6): 2155–2185.

Bound, John, David A Jaeger, and Regina M Baker. 1995. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American statistical association* 90 (430): 443–450.

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.

Chen, Jun, and Michael Ewens. 2021. *Venture Capital and Startup Agglomeration.* Technical report. National Bureau of Economic Research.

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.

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

Dalla Fontana, Silvia, Liudmila Alekseeva, Caroline Genc, and Hedieh Rashidi Ranjbar. 2025. "From In-Person to Online: The New Shape of the VC Industry." *Available at SSRN 5139704.*

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, and Matt Marx. 2024. "Firm Age and Invention: An Open-Access Dataset." *Working Paper.*

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. 2025. "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.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik instruments: What, when, why, and how." *American Economic Review* 110 (8): 2586–2624.

Gompers, 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, Anna Kovner, Josh Lerner, and David Scharfstein. 2010. "Performance persistence in entrepreneurship." *Journal of financial economics* 96 (1): 18–32.

Gompers, Paul, and Josh Lerner. 1999. *What drives venture capital fundraising?*

———. 2001. "The venture capital revolution." *Journal of economic perspectives* 15 (2): 145–168.

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.

Han, Pengfei, Chunrui Liu, Xuan Tian, and Kexin Wang. 2025. "Invest local or remote? The effects of COVID-19 lockdowns on venture capital investment around the world." *Management Science.*

Harris, Robert S, Tim Jenkinson, Steven N Kaplan, and Ruediger Stucke. 2023. "Has persistence persisted in private equity? Evidence from buyout and venture capital funds." *Journal of Corporate Finance* 81:102361.

Hegde, Deepak, and Justin Tumlinson. 2014. "Does social proximity enhance business partnerships? Theory and evidence from ethnicity's role in US venture capital." *Management Science* 60 (9): 2355–2380.

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 (4): 323–347.

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

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, and Antoinette Schoar. 2005. "Private equity performance: Returns, persistence, and capital flows." *The journal of finance* 60 (4): 1791–1823.

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

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, Josh, Antoinette Schoar, and Wan Wongsunwai. 2007. "Smart institutions, foolish choices: The limited partner performance puzzle." *The Journal of Finance* 62 (2): 731–764.

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.

Myers, Stewart C, and Nicholas S Majluf. 1984. "Corporate financing and investment decisions when firms have information that investors do not have." *Journal of financial economics* 13 (2): 187–221.

Nelson, Charles, and Richard Startz. 1988. *Some further results on the exact small sample properties of the instrumental variable estimator.*

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.

Raymond, Lindsey. 2023. *The market effects of algorithms.* Technical report. Working Paper.

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.

Stiglitz, Joseph E, and Andrew Weiss. 1981. "Credit rationing in markets with imperfect information." *The American economic review* 71 (3): 393–410.

Stock, James H, Jonathan H Wright, and Motohiro Yogo. 2002. "A survey of weak instruments and weak identification in generalized method of moments." *Journal of business & economic statistics* 20 (4): 518–529.

Stock, James H, and Motohiro Yogo. 2002. *Testing for weak instruments in linear IV regression.*

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.

Venugopal, Buvaneshwaran, and Vijay Yerramilli. 2025. "Nonexecutive Directors at Early-Stage Startups." *The Review of Corporate Finance Studies,* cfaf007.

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 2005 and 2024 the number and percentage of investments made by VC firms classified as data-driven.

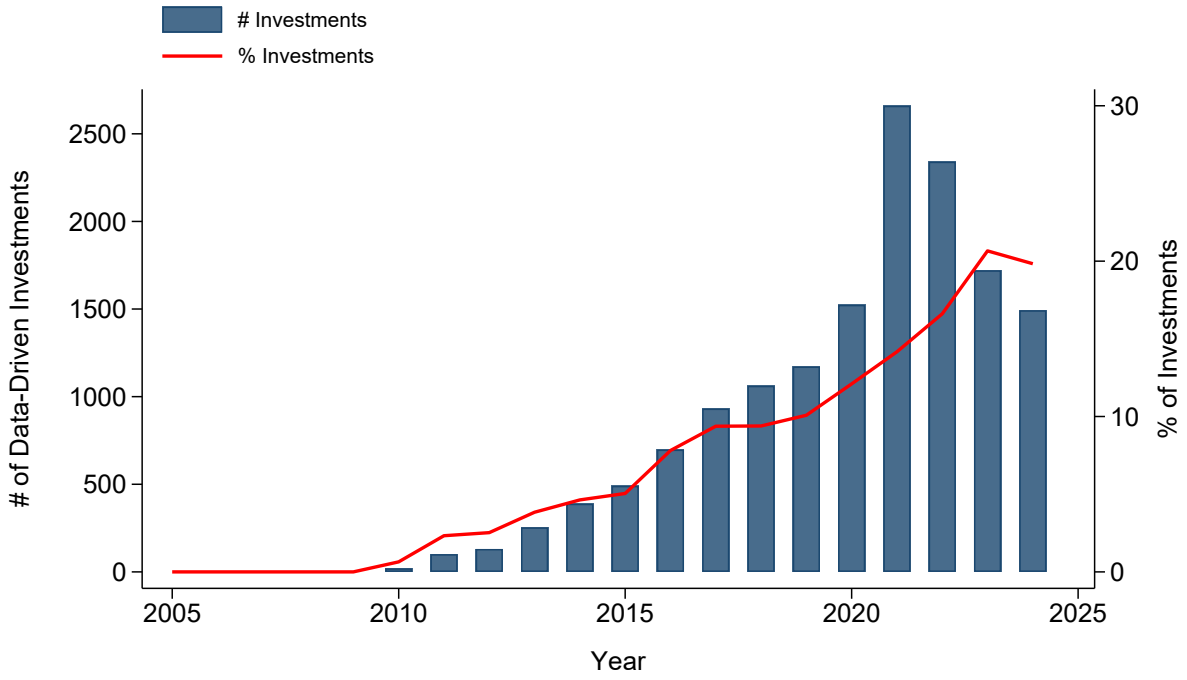


Figure 2: **Data Technology Adoption and Non-Hub Investments**

The figure plots the estimated coefficients from Equation 3 at the VC-Year level, of each year relative to VCs' adoption of data technologies. The dependent variable is the number of investments in non-hub commuting zones made by a VC firm in a given year. The horizontal bars represent the 95% 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.

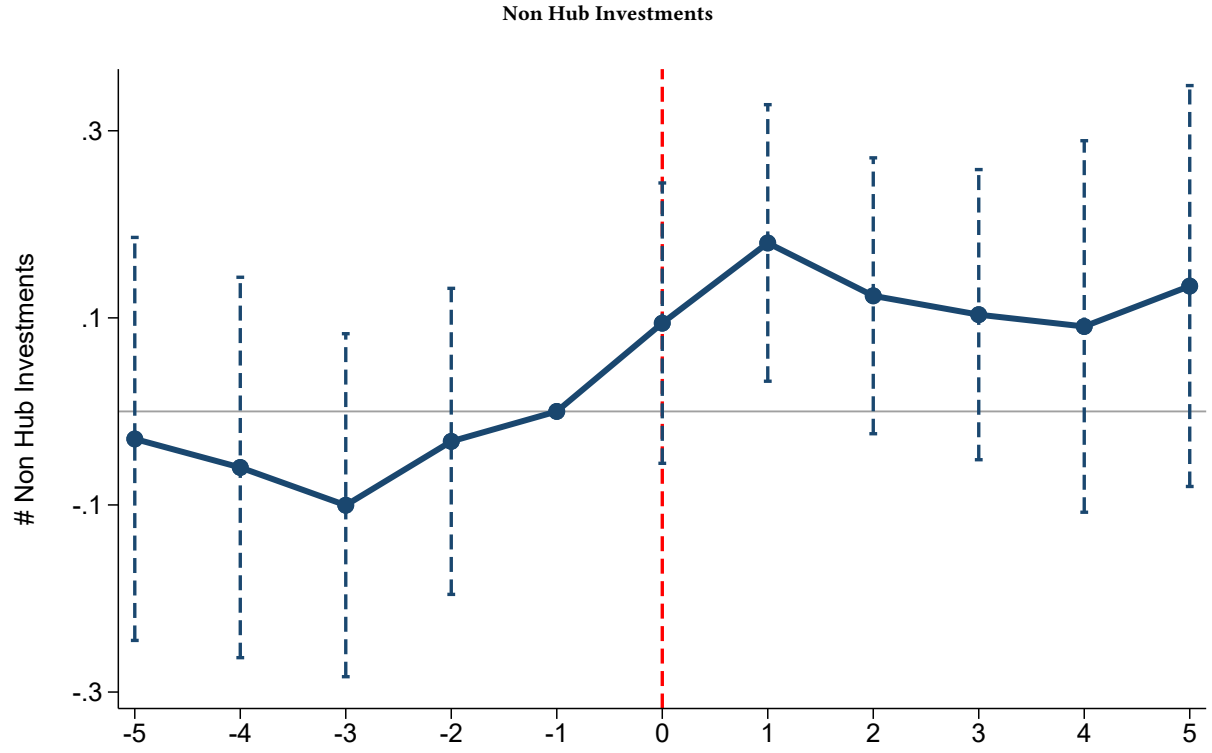
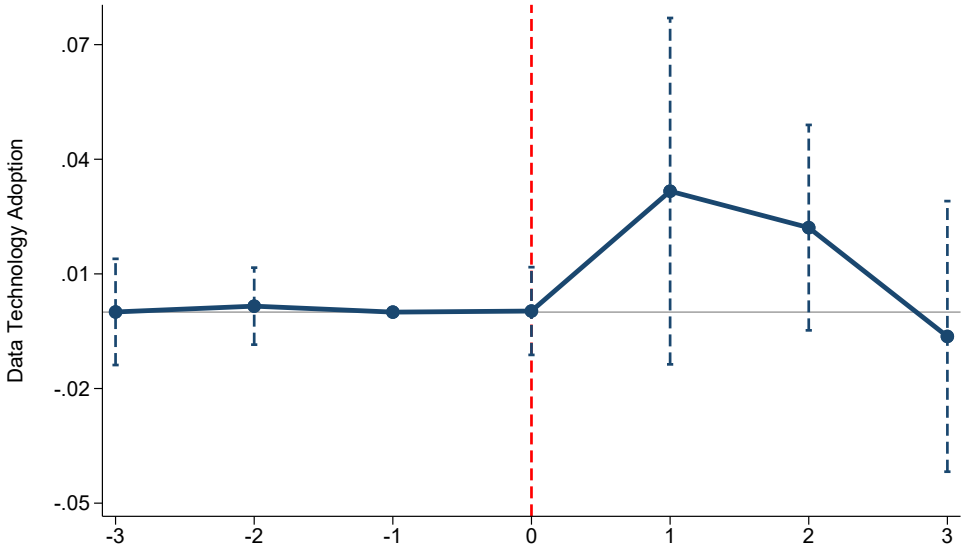


Figure 3: **Data Technologies And IV**

The figures plot the estimated coefficients from Equation 9 at the VC-Year level, of each year relative to VCs' raising a new fund. In Panel A, the dependent variable is an indicator equal to one if the VC hires a data scientist and zero otherwise. In Panel B, the dependent variable is the number of investments made in non-hub commuting zones by a VC firm in a given year. The year prior to VCs raising a new fund is the excluded category, reported as zero in the figures. The horizontal bars represent the 95% 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) Data Technology Adoption



(B) Non Hub Investments

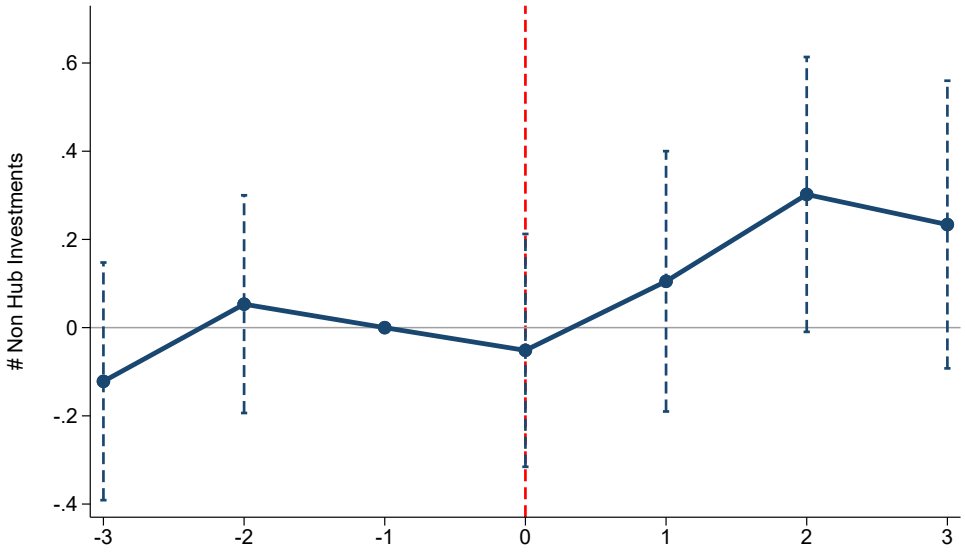


Table 1: **Summary Statistics**

	Mean	St. Dev.	P1	P25	Median	P75	P99	N
Panel A: Investment and Counterfactual Level								
Investment	0.13	34.29	0.00	0.00	0.00	0.00	1	1032103
Data Driven	0.10	0.31	0.00	0.00	0.00	0.00	1.00	1032103
Non Hub	0.33	0.47	0.00	0.00	0.00	1.00	1.00	1032103
Secondary Hub	0.21	0.41	0.00	0.00	0.00	0.00	1.00	1032103
Low Hub	0.13	0.33	0.00	0.00	0.00	0.00	1.00	1032103
Investor Age	14.37	14.89	0.00	5.00	10.00	18.00	96.00	1032103
AUM Prev 5 Years $(100,000s)$	5545	11942	0.00	273	1280	5029	77442	1032103
Number of Employees	15.36	29.55	0.00	3.00	6.00	15.00	214.00	1032103
Startup Age	3.15	2.62	0.00	1.00	3.00	5.00	12.00	1032103
Accelerator	0.32	0.47	0.00	0.00	0.00	1.00	1.00	1032103
Same Alma Mater	0.09	0.29	0.00	0.00	0.00	0.00	1.00	1032103
Non Top 20 University	0.56	0.50	0.00	0.00	1.00	1.00	1.00	1032103
First Time Founder	0.64	0.48	0.00	0.00	1.00	1.00	1.00	1032103
Panel B: VC-Year Level								
Data Driven	0.02	0.15	0.00	0.00	0.00	0.00	1.00	26180
# Investments	5.37	11.78	0.00	1.00	2.00	6.00	48.00	26180
# Hub Investments	3.37	7.70	0.00	0.00	1.00	3.00	38.00	26180
# Non Hub Investments	1.99	5.48	0.00	0.00	1.00	2.00	16.00	26180
# Low Hub Investments	0.12	0.57	0.00	0.00	0.00	0.00	2.00	26180
# Industries	5.37	3.78	0.00	1.00	2.00	6.00	12.00	26180
# Investments Outside Main Industry	4.75	7.49	0.00	1.00	2.00	5.00	42.00	20824
Investor Age	13.91	17.31	0.00	5.00	9.00	17.00	90.00	26180
AUM Prev 5 Years $(100,000s)$	1934	10827	10	49	266	1012	25897	26180
# of Employees	6.65	20.19	1.00	2.00	3.00	6.00	67.00	26180
Panel D: VC-Year Level - IV Sample								
Data Driven	0.03	0.17	0.00	0.00	0.00	0.00	1.00	5553
# Investments	10.53	19.88	1.00	2.00	5.00	11.00	78.00	5553
# Hub Investments	6.93	12.38	0.00	1.00	3.00	7.00	59.00	5553
# Non Hub Investments	3.60	9.96	0.00	1.00	2.00	4.00	22.00	5553
# Low Hub Investments	0.23	0.89	0.00	0.00	0.00	0.00	3.00	5553
# Industries	5.29	3.70	1.00	2.00	4.00	8.00	12.00	5553
# Investments Outside Main Industry	7.71	15.72	0.00	1.00	3.00	8.00	62.00	5553
Panel D: Investment Level								
Data Driven	0.11	0.31	0.00	0.00	0.00	0.00	1.00	140466
Follow On	0.58	0.49	0.00	0.00	1.00	1.00	1.00	140466
Major Exit	0.27	0.45	0.00	0.00	0.00	1.00	1.00	140466
IPO	0.06	0.23	0.00	0.00	0.00	0.00	1.00	140466
Acquisition	0.23	0.42	0.00	0.00	0.00	0.00	1.00	140466
Non Hub	0.37	0.48	0.00	0.00	0.00	1.00	1.00	140466
Different Industry	0.59	0.49	0.00	0.00	1.00	1.00	1.00	140466
Panel E: Commuting Zone - Year Level								
Data-Driven Entry	0.01	0.09	0.00	0.00	0.00	0.00	0.00	55715
# Funding Rounds	0.31	1.64	0.00	0.00	0.00	0.00	1.00	55715
# First VC Financing	0.15	0.78	0.00	0.00	0.00	0.00	0.00	55715
# First Investor	0.42	2.39	0.00	0.00	0.00	0.00	0.00	55715
# Unique Investors	0.57	3.23	0.00	0.00	0.00	0.00	0.00	55715
# VC Patents	0.33	1.66	0.00	0.00	0.00	0.00	1.00	55715
# Investments t-5 to t-1	0.13	0.71	0.00	0.00	0.00	0.00	0.00	55605
GDP mil	2.03	3.25	0.08	0.50	0.97	2.30	4.75	55715
Income	37519	9264	24690	31394	35744	41221	48374	55715
Percentage College	0.19	0.06	0.10	0.16	0.18	0.22	0.26	53625
RECPI	0.47	1.88	0.00	0.02	0.06	0.27	1.04	55605

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
Age	15.29	566	13.88	25,614	1.4**
# Employees	10.72	566	6.20	25,614	4.52***
AUM ($ Mil)	11,691	566	1,718	25,614	9,972***
# Total Investments	26.48	566	4.90	25,614	21.58***
# Hub	18.35	566	3.04	25,614	15.31***
# Non Hub	8.13	566	1.86	25,614	6.28***
# Low Hub	0.32	566	0.12	25,614	0.21***

Table 3: **Do Data-Driven VCs Tilt their Portfolios Toward Startups Located in Non Hubs?**
This table reports results for regressions for the limited probability model in Equation 1. The dependent variable is an indicator equal to one (multiplied by 100) for actual matched startup-VC pairs, and zero for counterfactual startup-VC pairs. All specifications include group fixed effects. Standard errors are clustered at the VC-Firm level.

Outcomes:	1(Investment)×100			
	(1)	(2)	(3)	(4)
Data Driven× Non Hub	0.466*	0.465*	0.468*	0.533**
	(1.66)	(1.66)	(1.67)	(1.99)
Non Hub	-1.144***	-1.149***	-1.083***	-1.008***
	(-15.33)	(-15.33)	(-14.41)	(-13.40)
Log(Investor Age)		-0.458***	-0.465***	-0.419***
		(-9.76)	(-9.90)	(-8.92)
Log(Investor AUM)		0.173***	0.174***	0.173***
		(15.76)	(15.83)	(15.76)
Log(Investor Employees)		0.734***	0.738***	0.573***
		(19.40)	(19.47)	(14.57)
Log(Startup Age)			-0.710***	-0.763***
			(-11.64)	(-12.49)
1(Accelerator)			0.534***	0.525***
			(7.97)	(7.83)
Same Alma Mater				2.106***
				(15.00)
Data-Driven	Yes	Yes	Yes	Yes
Group FE	Yes	Yes	Yes	Yes
R-squared	0.11	0.12	0.11	0.13
N	802,008	802,008	802,008	802,008

Table 4: **Do Data-Driven VCs Tilt their Portfolios Toward Startups Located in Non Hubs?**

This table reports results for regressions for the limited probability model in Equation 1. The dependent variable is an indicator equal to one (multiplied by 100) for actual matched startup-VC pairs, and zero for counterfactual startup-VC pairs. All specifications include group fixed effects. Standard errors are clustered at the VC-Firm level.

Outcomes:	1(Investment)×100					
	(1)	(2)	(3)	(4)	(5)	(6)
Data Driven×Low Hub	0.640	0.742*				
	(1.51)	(1.74)				
Data Driven×Secondary Hub	0.369	0.415				
	(1.11)	(1.25)				
Data Driven×Non Hub×Non Top20			1.034*	0.985*		
			(1.78)	(1.70)		
Data Driven×Non Hub×First-Time Founder					0.380	0.376
					(0.65)	(0.64)
Data Driven×Non Hub			-0.215	-0.162	0.251	0.317
			(-0.47)	(-0.35)	(0.53)	(0.67)
Log(Investor Age)	-0.465***	-0.419***	-0.466***	-0.421***	-0.465***	-0.419***
	(-9.90)	(-8.92)	(-9.91)	(-8.95)	(-9.89)	(-8.92)
Log(Investor AUM)	0.174***	0.173***	0.175***	0.174***	0.174***	0.173***
	(15.83)	(15.76)	(15.88)	(15.79)	(15.85)	(15.78)
Log(Investor Employees)	0.738***	0.574***	0.739***	0.575***	0.737***	0.575***
	(19.47)	(14.57)	(19.48)	(14.55)	(19.45)	(14.60)
Log(Startup Age)	-0.708***	-0.761***	-0.696***	-0.758***	-0.680***	-0.739***
	(-11.59)	(-12.46)	(-11.40)	(-12.39)	(-11.06)	(-11.99)
1(Accelerator)	0.534***	0.524***	0.520***	0.520***	0.546***	0.534***
	(7.96)	(7.82)	(7.76)	(7.76)	(8.14)	(7.97)
Same Alma Mater		2.106***		2.090***		2.081***
		(14.99)		(14.11)		(14.80)
Data-Driven	Yes	Yes	Yes	Yes	Yes	Yes
Non Hub	No	No	Yes	Yes	Yes	Yes
Secondary Hub	Yes	Yes	No	No	No	No
Low Hub	Yes	Yes	No	No	No	No
Group FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.11	0.12	0.11	0.13	0.11	0.13
N	802,008	802,008	802,008	802,008	802,008	802,008

Table 5: **Do Data-Driven VCs Increase Their Investments in Non Hub Startups?**

This table reports results for regressions for the model in Equation 4. The dependent variable is the number of investments in non-hub 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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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.

Outcomes:	# Investments			
	(1)	(2)	(3)	(4)
Data Driven	0.336***	0.183**	0.135*	
	(4.02)	(2.31)	(1.85)	
Data Driven × Non Hub			0.137*	0.137*
			(1.86)	(1.86)
Non Hub			-1.150***	-1.150***
			(-58.23)	(-58.23)
Log(Investor Age)	0.539***	-0.060	-0.116*	
	(9.32)	(-0.80)	(-1.91)	
Log(Investor AUM)		0.053***	0.052***	
		(8.19)	(11.08)	
Log(Investor Employees)		0.517***	0.499***	
		(8.05)	(9.42)	
VC-Firm FE	Yes	Yes	Yes	No
VC-Firm×Year FE	No	No	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
R-squared	0.11	0.12	0.11	0.13
N	13,209	13,209	26,418	26,418

Table 6: **Do Data-Driven VCs Increase Their Investments Outside Their Industry?**

This table reports results for regressions for the model in Equation 4. In column (1), the dependent variable in the number of industries the VC invests in outside their focal industry. In column (2), the dependent variable is the number of investments made in industries outside of their focal industry. In column (3), the dependent variable is the number of investments made in the Biotechnology and Health Care industries. In column (4), the dependent variable is the number of investments in the Energy and Sustainability industry. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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.

Outcomes:	# Industries		# Investments		# Health Care		# Energy & Manuf.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Data Driven	0.210***	0.072**	0.359***	0.126*	0.409***	0.316***	0.104	0.176**
	(3.03)	(2.58)	(2.85)	(1.74)	(3.05)	(3.73)	(0.64)	(2.52)
Log(Investor Age)	-0.116***	-0.068*	-0.189***	-0.198***	-0.147**	-0.155**	-0.219***	-0.221**
	(-9.13)	(-1.76)	(-6.25)	(-5.40)	(-2.49)	(-1.97)	(-3.48)	(-2.47)
Log(Investor AUM)	0.091***	0.031***	0.197***	0.062***	0.283***	0.104***	0.208***	0.085***
	(7.28)	(10.30)	(6.86)	(9.18)	(8.05)	(5.85)	(5.15)	(5.62)
Log(Investor Employees)	0.266***	0.345***	0.392***	0.468***	0.245***	0.390***	0.469***	0.491***
	(14.00)	(18.01)	(11.02)	(21.34)	(4.73)	(8.11)	(5.98)	(7.69)
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	0.11	0.12	0.11	0.13				
N	18,742	18,742	18,742	18,742	18,742	18,742	18,742	18,742

Table 7: **Do Data-Driven VCs Increase Their Investments in Non Hub Startups - Reduced Form?**
This table reports results for regressions for the model in Equation 8. In column (1), the dependent variable is an indicator if the VC hires a data scientist. In columns (2) through (4), the dependent variable is the number of investments in non-hub 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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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 an OLS regression in column (1) and a Poisson Pseudo Maximum Likelihood (PPML) regression in columns (2) through (4). Standard errors are clustered at the VC-Firm level and bootstrapped over 1000 iterations for columns (2) through (4).

Outcomes:	Adoption	# Investments		
	(1)	(2)	(3)	(4)
VC Exposure×Post	0.059***	0.261***	0.189***	
	(4.45)	(2.65)	(2.66)	
VC Exposure×Post×Non Hub			0.107**	0.108***
			(2.02)	(2.59)
Non Hub	No	No	Yes	Yes
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	No
VC-Firm×Year FE	No	No	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
F-statistic	22.23			
N	3582	3166	6332	6332
R-squared	-0.17	0.56	0.68	0.73

Table 8: **Do Data-Driven VCs Increase Their Investments Outside Their Industry - Reduced Form?**

This table reports results for regressions for the model in Equation 8. In column (1), the dependent variable is an indicator if the VC hires a data scientist. In column (3), the dependent variable is the number of investments made in industries outside of their focal industry. In column (4), the dependent variable is the number of investments made in the Biotechnology and Health Care industries. In column (5), the dependent variable is the number of investments in the Energy and Sustainability industry. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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 an OLS regression in column (1) and a Poisson Pseudo Maximum Likelihood (PPML) regression in columns (2) through (5). Standard errors are clustered at the VC-Firm level and bootstrapped over 1000 iterations for columns (2) through (4).

Outcomes:	Adoption	# Industries	# Investments	# Health Care	# Energy & Manuf.
	(1)	(2)	(3)	(4)	(5)
VC Exposure×Post	0.046***	0.140*	0.172**	0.444**	0.378***
	(3.59)	(1.87)	(2.36)	(2.33)	(3.60)
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes
F-statistic	19.4				
N	4039	4039	4039	4039	4039
R-squared	-0.17	0.41	0.77	0.69	0.61

Table 9: **Data Technology Adoption and Follow On Funding**

This table reports results for regressions for the model in Equation 11. The dependent variable is an indicator equal to one if the startup received another round of financing. "State" denotes the state where the startup is headquartered. "Industry" denotes the startup's industry (among twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "Stage" denotes the stage of investment (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.

Outcomes:	Follow On				
	(1)	(2)	(3)	(4)	(5)
Data Driven	0.013*	0.010		0.012	
	(1.70)	(1.41)		(0.97)	
Data Driven×Non Hub		-0.013*	-0.013***		
		(-1.68)	(-2.39)		
Data Driven×Diff Industry				0.001	0.001
				(0.08)	(0.12)
Non Hub		-0.013***	-0.009***		
		(-7.53)	(-5.67)		
Diff Industry				-0.010**	-0.010**
				(-2.37)	(-2.30)
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	No	Yes	No
VC-Firm FE×Year	No	No	Yes	No	Yes
State×Industry×Stage×Year	Yes	Yes	Yes	Yes	Yes
Data Driven×Non Hub = Data Driven (p-value)		0.08*			
Data Driven×Non Hub = Non Hub (p-value)		0.99	0.26		
Data Driven×Diff Industry = Data Driven (p-value)				0.50	
Data Driven×Diff Industry = Diff Industry (p-value)				0.63	0.51
N	109359	71738	71738	109359	109359
R-squared	0.29	0.27	0.27	0.29	0.29

Table 10: **Data Technology Adoption and Major Exit**

This table reports results for regressions for the model in Equation 11. The dependent variable is an indicator equal to one if the startup achieved a major exit outcome, either through a high-quality acquisition or IPO. "State" denotes the state where the startup is headquartered. "Industry" denotes the startup's industry (among twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "Stage" denotes the stage of investment (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.

Outcomes:	Major Exit				
	(1)	(2)	(3)	(4)	(5)
Data Driven	-0.029***	-0.027**		-0.037***	
	(-3.67)	(-2.03)		(-4.52)	
Data Driven×Non Hub		0.019*	0.014		
		(1.73)	(1.62)		
Data Driven×Diff Industry				0.011**	0.011*
				(2.19)	(2.07)
Non Hub		-0.065***	-0.065***		
		(-16.47)	(-12.81)		
Diff Industry				-0.010***	-0.013***
				(-6.07)	(-4.29)
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	No	Yes	No
VC-Firm FE×Year	No	No	Yes	No	Yes
State×Industry×Stage×Year	Yes	Yes	Yes	Yes	Yes
Data Driven×Non Hub = Data Driven (p-value)		0.09*			
Data Driven×Non Hub = Non Hub (p-value)		0.00***	0.00***		
Data Driven×Diff Industry = Data Driven (p-value)				0.03**	
Data Driven×Diff Industry = Diff Industry (p-value)				0.00***	0.05**
N	66805	37929	37929	66805	66805
R-squared	0.43	0.36	0.35	0.43	0.42

Table 11: **Data-Driven Investment Entry and VC Activity**

This table reports results for regressions for the model in Equation 11. 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. All columns include cohort by year fixed effects and cohort by commuting zone fixed effects. I estimate a Poisson Pseudo Maximum Likelihood (PPML) regression. Standard errors 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

A Robustness

Table A1: **Do Data-Driven VCs Tilt their Portfolios Toward Startups Located in Non Hubs - First Time Investments?**

This table reports results for regressions for the limited probability model in Equation 1. The dependent variable is an indicator equal to one (multiplied by 100) if actual matched startup-VC pairs, and zero for counterfactual startup-VC pairs. All specifications include group fixed effects. Standard errors are clustered at the VC-Firm level.

Outcomes:	1(Investment)×100			
	(1)	(2)	(3)	(4)
Data Driven×Non Hub	0.806**		0.465	0.292
	(2.30)		(0.77)	(0.48)
Data Driven×Low Hub		0.796*		
		(1.81)		
Data Driven×Secondary Hub		0.612*		
		(1.95)		
Data Driven×Non Hub×Non Top20			0.381**	
			(2.51)	
Data Driven×Non Hub×First-Time Founder				0.817
				(1.10)
Data-Driven	Yes	Yes	Yes	Yes
Non Hub	Yes	No	Yes	Yes
Secondary Hub	No	Yes	No	No
Low Hub	No	Yes	No	No
Non Top20	No	No	Yes	No
First-Time Founder	No	No	No	Yes
Group FE	Yes	Yes	Yes	Yes
Group FE	Yes	Yes	Yes	Yes
N	529486	529486	529486	529486
R-Squared	-0.11	-0.11	-0.11	-0.11

Table A2: **Do Data-Driven VCs Increase Their Investments Outside Their Network - First Time Investments?**

This table reports results for regressions for the model in Equation 4. In column (1), the dependent variable is the number of investments in non-hub commuting zones. In column (2), the dependent variable is the number of investments made in industries outside of their focal industry. In column (3), the dependent variable is the number of investments made in the Biotechnology and Health Care industries. In column (4), the dependent variable is the number of investments in the Energy and Manufacturing industry. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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.

Outcomes:	# Non Hub	# Investments	# Health Care	# Energy & Manuf.
	(1)	(2)	(3)	(4)
Data Driven	0.109**	0.119**	0.324***	0.174
	(2.21)	(2.26)	(2.97)	(1.53)
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
N	13209	17184	17184	17184
R-squared	0.48	0.59	0.6	0.53

Table A3: **Do Data-Driven VCs Increase Their Investments in Non Hub Startups - Alternative Definitions?**

This table reports results for regressions for the model in Equation 4. In column (1), the dependent variable in the number of non-hub commuting zones the VC invests in. In column (2), the dependent variable is the number of investments made in secondary hubs. In column (3), the dependent variable is the number of investments made non-secondary hubs. In column (4), the dependent variable is the number of investments made in low activity hubs. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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.

Outcomes:	# Commuting Zones	# Secondary Hubs	# Non Secondary Hubs	# Low Activity
	(1)	(2)	(3)	(4)
Data Driven	0.094*	0.211**	0.143*	0.672***
	(1.87)	(2.48)	(2.13)	(2.68)
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
N	13209	13209	13209	13209
R-squared	0.39	0.52	0.43	0.48

Table A4: **Do Data-Driven VCs Increase Their Investments Outside Their Networks Before 2020?**

This table reports results for regressions for the model in Equation 4. The sample is restricted to end in 2019. In column (1), the dependent variable is the number of investments in non-hub commuting zones. In column (2), the dependent variable is the number industries the VC invests in outside their focal industry. In column (3), the dependent variable is the number of investments made in industries outside of their focal industry. In column (4), the dependent variable is the number of investments made in the Biotechnology and Health Care industries. In column (5), the dependent variable is the number of investments in the Energy and Manufacturing industry. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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.

Outcomes:	# Non Hub	# Industries	# Investments	# Health Care	# Energy & Manuf.
	(1)	(2)	(3)	(4)	(5)
Data Driven	0.095**	0.044**	0.096**	0.241	0.178***
	(2.21)	(2.06)	(2.36)	(1.29)	(3.55)
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes
N	8296	10999	10999	10999	10999
R-squared	0.47	0.39	0.66	0.65	0.59

Table A5: **Do Data-Driven VCs Increase Their Investments Outside Their Industry?**

This table reports results for regressions for the model in Equation 4. The sample is restricted to end in 2019. In column (1), the dependent variable is the number of investments in non-hub commuting zones. In column (2), the dependent variable is the number industries the VC invests in outside their focal industry. In column (3), the dependent variable is the number of investments made in industries outside of their focal industry. In column (4), the dependent variable is the number of investments made in the Biotechnology and Health Care industries. In column (5), the dependent variable is the number of investments in the Energy and Manufacturing industry. In Panel A, the main explanatory variable is the natural log of one plus the number of data-related employees. In Panel B, the main explanatory variable is the number of data-related employees per partner. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Consumers, Financial Services, Health Care, Internet Services, Manufacturing and Sustainability, Media and Entertainment, Mobiles and Apps, Professional Services, and Software). "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.

Outcomes:	# Non Hub	# Industries	# Investments	# Health Care	# Energy & Manuf.
	(1)	(2)	(3)	(4)	(5)
Panel A: Data Driven Intensity					
Log(Data Employees)	0.170**	0.091***	0.161***	0.349***	0.158
	(2.52)	(2.88)	(3.06)	(3.30)	(1.28)
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes
N	13209	18985	18985	18985	18985
R-squared	0.4	0.36	0.63	0.65	0.57
Panel B: Data Scientist per Investment Partner					
Data Employee / Partners	0.120***	0.056***	0.110***	0.149***	0.081***
	(5.18)	(3.21)	(4.27)	(3.01)	(3.16)
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes
N	13209	18985	18985	18985	18985
R-squared	0.4	0.36	0.63	0.65	0.57

B Instrumental Variables

Table B1: **Crunchbase AI Categories**

Artificial Intelligence (AI)	Generative AI
Augmented Reality	Geospatial
Autonomous Vehicles	Human Computer Interaction
Big Data	Image Recognition
Business Intelligence	Machine Learning
Chatbot	Natural Language Processing
Computer Vision	Predictive Analytics
Data Collection and Labeling	Quantum Computing
Data Mining	Robotic Process Automation
Data Visualization	Robotics
Drone Management	Speech Recognition
Embedded Systems	Text Analytics
Facial Recognition	Virtual Reality

Table B2: **Is Hiring Associated with Raising a New Fund?**

This table reports results for regressions for the model in Equation 4. In column (1) through (3), the dependent variable is the natural log of the number of employees at a VC firm in a given year. In columns (4) through (6), the dependent variable is the percentage change in the number of employees from the previous year. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Cybersecurity, Energy and Sustainability, Financial Services, Health Care, Internet Services, Media, Professional Services, Semiconductors, and Software). "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 an OLS regression. Standard errors are clustered at the VC-Firm level.

Outcomes:	Log(Num Employees)			$\%\Delta$ Employees		
	(1)	(2)	(3)	(4)	(5)	(6)
Raise Fund$_t$	0.067***			0.043***		
	(7.28)			(5.08)		
Raise Fund$_{t-2:t}$		0.122***			0.016**	
		(9.21)			(2.20)	
Raise Fund$_{t+1}$			0.023**			-0.000
			(2.35)			(-0.03)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
State$\times$Industry$\times$Stage$\times$Year	Yes	Yes	Yes	Yes	Yes	Yes
N	16524	16524	15077	15355	15355	13914
R-sqaured	0.87	0.88	0.87	0.11	0.10	0.10

Table B3: **Correlation of Fundraising Across VC Firms (2010–2024)**

Panel A shows correlations of raising a new fund between 10 randomly selected Silicon Valley VC firms in the information technology sector. Panel B shows correlations of raising a new fund between 10 randomly selected Boston VC firms in the biotechnology sector.

Panel A										
	SVB	Icon	Translink	Ulu	DCVC	AMV	Triangle Peak	Zeev	True	Innovation Endeavors
SVB	1.00									
Icon	0.45	1.00								
Translink	0.44	0.53	1.00							
Ulu	-0.30	-0.25	-0.13	1.00						
DCVC	0.21	0.35	0.38	0.00	1.00					
AMV	-0.30	-0.25	-0.13	0.17	-0.35	1.00				
Triangle Peak	-0.24	-0.20	-0.10	-0.20	0.14	0.29	1.00			
Zeev	0.44	-0.13	-0.07	-0.13	-0.19	-0.13	-0.10	1.00		
True	0.34	0.20	0.29	0.20	0.19	0.20	0.03	0.29	1.00	
Innovation Endeavors	-0.02	0.08	0.44	0.08	0.21	-0.30	0.21	-0.16	-0.26	1.00
Average Correlation	**-0.084**									
Panel B										
	PureTech Health	Omega	Third Rock	Atlas	Longwood	SR One	Advent Intl	MassVentures	Sanofi	Flagship Pioneering
PureTech Health	1.00									
Omega Funds	0.08	1.00								
Third Rock Ventures	-0.28	0.29	1.00							
Atlas Venture	-0.03	0.22	0.09	1.00						
Longwood Fund	-0.20	0.61	0.35	-0.20	1.00					
SR One	-0.10	0.33	-0.19	-0.29	0.53	1.00				
Advent International	0.08	0.17	0.58	-0.05	0.27	-0.22	1.00			
MassVentures	-0.10	-0.22	-0.19	-0.29	-0.13	-0.07	-0.22	1.00		
Sanofi Ventures	-0.10	0.33	0.38	-0.29	0.53	-0.07	0.33	-0.07	1.00	
Flagship Pioneering	0.03	0.05	-0.09	0.34	-0.13	-0.25	0.05	0.29	-0.25	1.00
Average Correlation	**-0.080**									

Table B4: **Do Data-Driven VCs Increase Their Investments in Non Hub Startups - Second Stage?**
This table reports results for regressions for the model in Equation 8 as well as the second stage IV results. In column (1) and (2) through (5), the dependent variable is the number of investments in non-hub locations. In column (2), the dependent variable is an indicator if the VC hires a data scientist. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Cybersecurity, Energy and Sustainability, Financial Services, Health Care, Internet Services, Media, Professional Services, Semiconductors, and Software). "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 an OLS regression in column (1) and a Poisson Pseudo Maximum Likelihood (PPML) regression in columns (2) through (4). Standard errors are clustered at the VC-Firm level.

Outcomes:	OLS	First Stage	2SLS		
	(1)	(2)	(3)	(4)	(5)
Adopt	0.161		4.314***	2.835***	
	(1.62)		(2.59)	(4.16)	
Adopt×Non Hub				0.154	0.154
				(0.90)	(1.17)
VC Exposure×Post		0.059***			
		(4.91)			
Non Hub	No	No	No	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	No
VC-Firm×Year FE	No	No	No	No	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	
N	3166	3166	3166	6332	6332
R-sqaured	0.56	-0.23	0.52	0.68	0.73

Table B5: **Do Data-Driven VCs Increase Their Investments in Non Hub Startups - Alternative Timing**

This table reports results for regressions for the model in Equation 8 as well as the second stage IV results. In column (1), $VCExposure \times Post$ is lagged by one. In column (2), $VCExposure \times Post$ is lagged by two. In column (3), $VCExposure \times Post$ is lagged by three. In column (4), $VCExposure \times Post$ is lagged by four. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Cybersecurity, Energy and Sustainability, Financial Services, Health Care, Internet Services, Media, Professional Services, Semiconductors, and Software). "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 an OLS regression in column (1) and a Poisson Pseudo Maximum Likelihood (PPML) regression in columns (2) through (4). Standard errors are clustered at the VC-Firm level.

Outcomes:	# Investments			
	$t+1$	$t+2$	$t+3$	$t+4$
	(1)	(2)	(3)	(4)
VC Exposure×Post	0.235***	0.114*	0.140**	0.143*
	(3.38)	(1.67)	(2.00)	(1.91)
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	No
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
N	2967	2758	2555	2359
R-sqaured	0.53	0.52	0.53	0.53

Table B6: **Do Data-Driven VCs Increase Their Investments Outside Their Main Industry - Second Stage?**

In column (1), the dependent variable is an indicator if the VC hires a data scientist. In column (3), the dependent variable is the number of investments made in industries outside of their focal industry. In column (4), the dependent variable is the number of investments made in the Biotechnology and Health Care industries. In column (5), the dependent variable is the number of investments in the Energy and Sustainability industry. "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 twelve industries: Artificial Intelligence, Biotechnology, Commerce, Cybersecurity, Energy and Sustainability, Financial Services, Health Care, Internet Services, Media, Professional Services, Semiconductors, and Software). "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 an OLS regression in column (1) and a Poisson Pseudo Maximum Likelihood (PPML) regression in columns (2) through (5). Standard errors are clustered at the VC-Firm level.

Outcomes:	Adopt	# Industries	# Investments	# Health Care	#Energy & Manuf.
	(1)	(2)	(3)	(4)	(5)
Adopt		1.152	3.812***	2.714**	8.135***
		(1.06)	(3.88)	(2.32)	(3.93)
VC Exposure×Post	0.046***				
	(4.36)				
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes
N	4039	4039	4039	4039	4039
R-sqaured	0.29	0.28	0.77	0.69	0.61

B.1 Discussion About LATE

In the presence of heterogeneous treatment effects, the instrument identifies a local average treatment effect (LATE), which reflects the causal effect on the subset of VCs whose adoption of data technologies is induced by the instrument (the "compliers"; Angrist and Pischke (2009)). Specifically, the estimates capture the causal effect for marginal VCs that are more likely to adopt data technologies when they both had higher pre-2010 exposure to AI and experience a new fundraising event, but would not have adopted absent this combination.

It is important to distinguish this complier group from "always-takers" (VCs that would adopt data technologies regardless of their AI exposure or fundraising cycle) and from "never-takers" (VCs that never adopt the technology). If VCs that are more predisposed to adopt (the always-takers) are also those with greater resources or stronger networks, then the LATE may overstate the effect relative to the average treatment effect (ATE). This is because such VCs likely would have expanded their investment activity even without data technologies. Conversely, the never-takers may be systematically different, perhaps smaller or more specialized funds, so that their potential response to adoption cannot be inferred from the complier group.

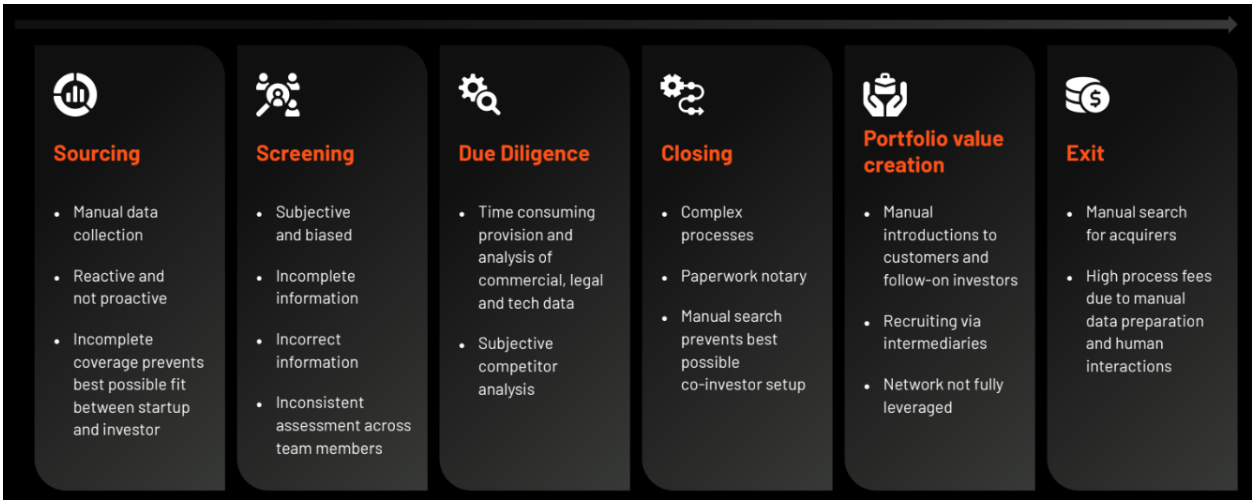
For intuition, consider that a top-tier firm such as Sequoia or Andreessen Horowitz may adopt data technologies regardless of prior exposure or fundraising cycles, while a small, highly specialized biotech VC may never adopt. By contrast, a mid-sized VC with some AI investments before 2010 might only choose to hire a data scientist when raising a new fund, making them a complier in this setting.

Nonetheless, the LATE is the most policy-relevant parameter in this context, since it identifies the causal effect for those VCs whose adoption is sensitive to plausibly exogenous shifts in exposure and fundraising timing. These are precisely the VCs at the margin of adoption - those for whom organizational frictions or resource constraints make the decision to adopt uncertain. Understanding their behavior provides the clearest counterfactual for how data technologies expand the geography of venture capital investment when adoption is not universal.

C Other

Figure C1: **Traditional VC Model**

(A) VC Investment Process
Source: Data Driven VC Landscape 2025



(B) Passion Capital Investments
Source: Tech Crunch: Passion Capital Proves Its Mettle As A Go-To Seed VC In Europe

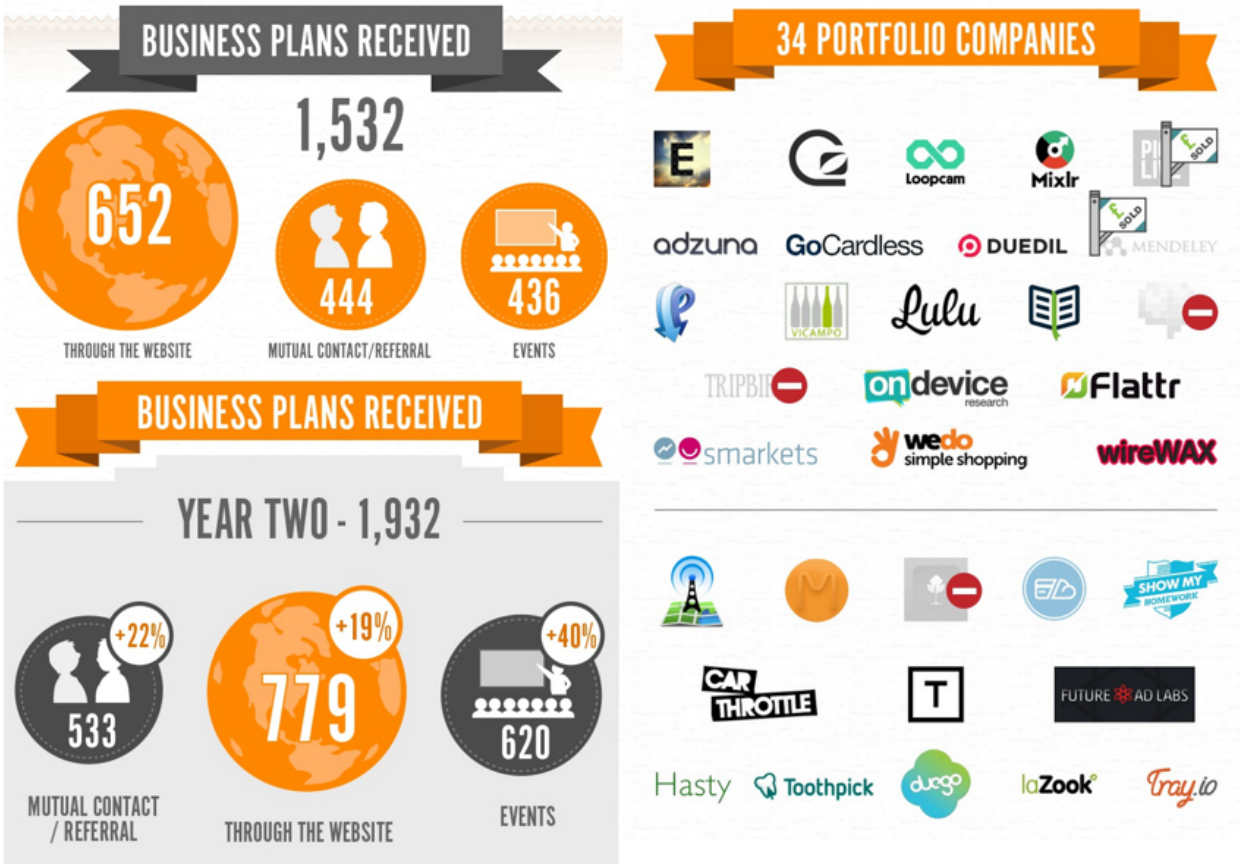
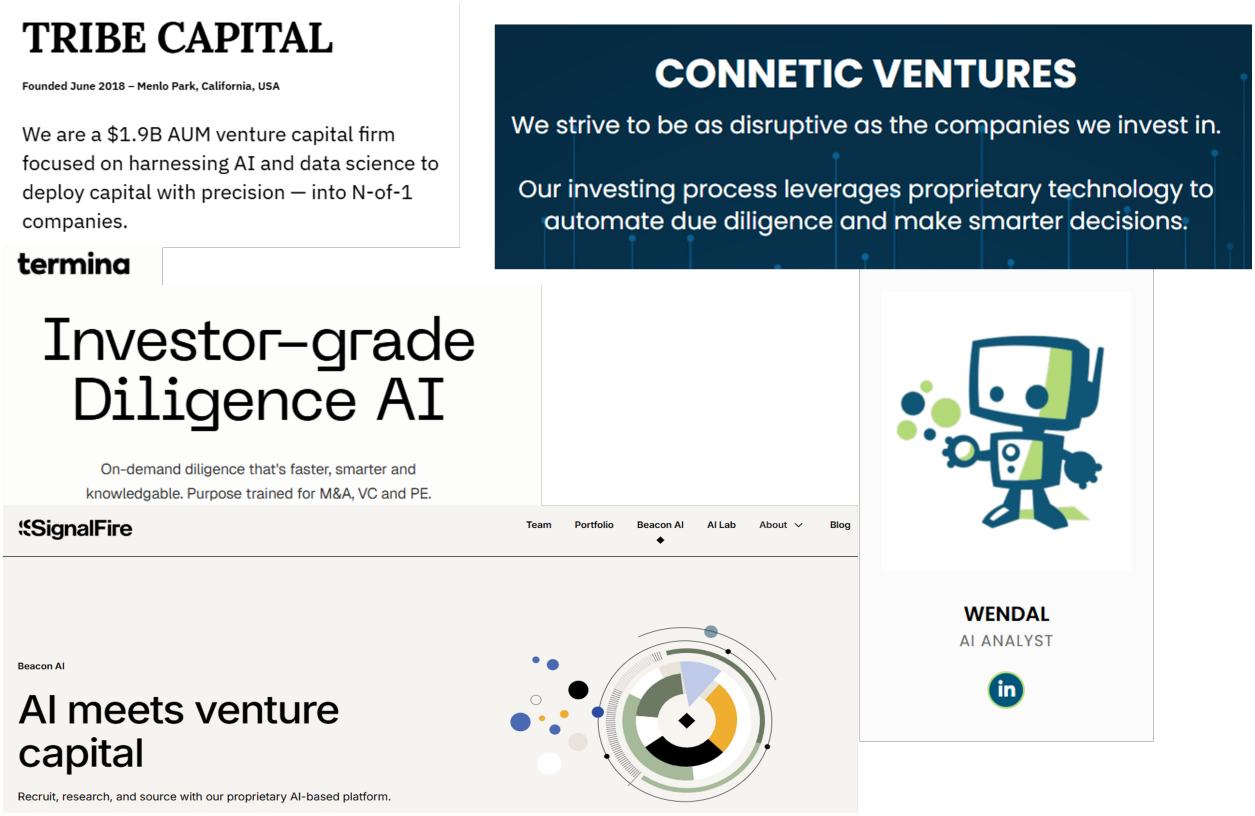


Figure C2: **Data Driven VC Model**

(A) Examples of Data-Driven VCs
Tribe Capital's Termina, SignalFire's Beacon AI, and Connetic Ventures' Wendel.



(B) Data Driven Use Cases
Source: Data Driven VC Landscape 2025

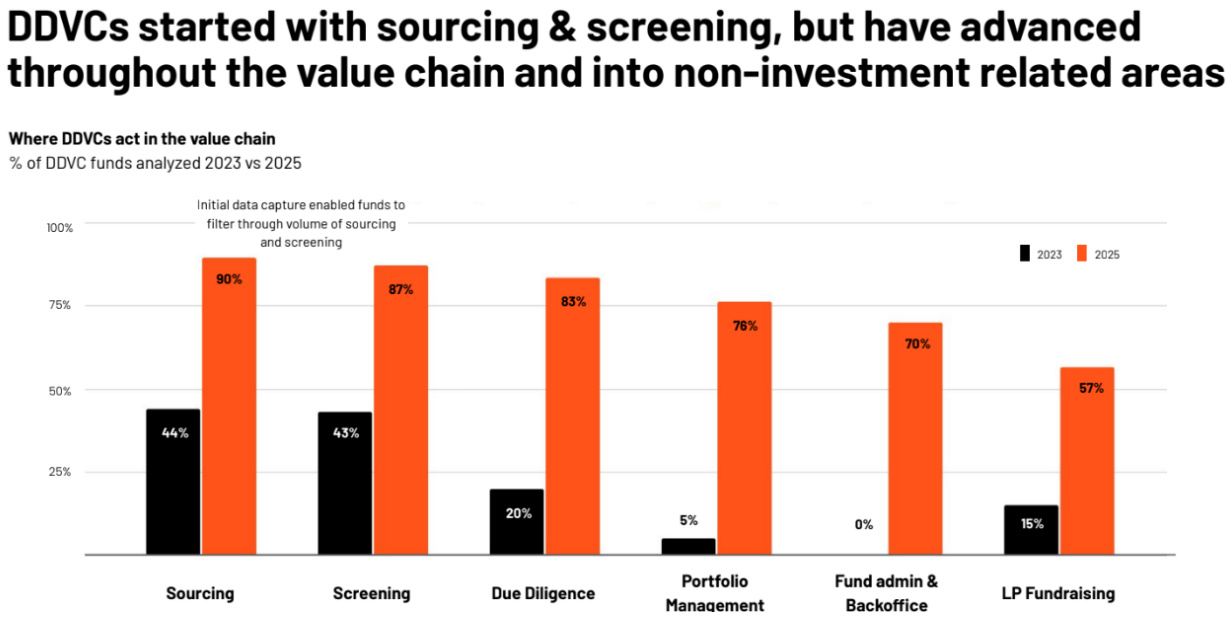
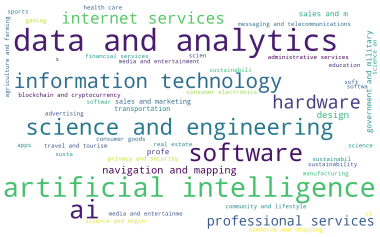
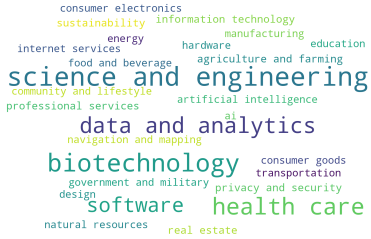


Figure C3: **Industry Classifications**

(A) Artificial Intelligence



(B) Biotechnology



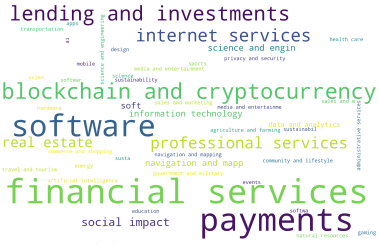
(C) Commerce



(D) Consumers



(E) Financial Services



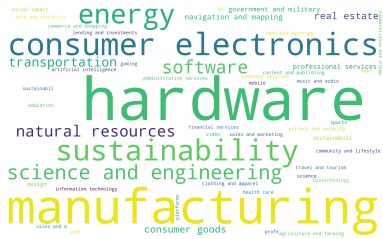
(F) Health Care



(G) Internet Services



(H) Manufacturing and Sustainability



(I) Media and Entertainment



(J) Mobile and Apps



(K) Professional Services



(L) Software

