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

Melissa Crumling[*]

Drexel University

July 31, 2024

Abstract

Financial intermediaries are increasingly using digital data and machine learning techniques to inform their investment decision process. One underexplored advantage of data-driven approaches is the use of these technologies to expand investors' investment opportunity sets. This paper examines the impact of data technologies on sourcing deal flow in the Venture Capital (VC) industry. The VC industry is critical for financing young, innovative firms, yet traditional deal sourcing methods are often limited to established networks and geographic clusters. I find that after VCs adopt data technologies, they are more likely to invest in firms outside of their typical networks, as proxied by startups located in areas with low history of VC activity. Results are robust to isolating plausibly exogenous variations in VCs' pre-exposure to data technologies, specifically artificial intelligence, suggesting a causality between data technology adoption and these effects. In addition, I find that data-driven investments in areas with historically low VC presence stimulate entrepreneurial activity, suggesting potential policy implications for fostering regional 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 Jason (Pang-Li) Chen, Winston Dou (Rising Scholars mentor), Tanja Kirmse, Torin McFarland, and participants of the Rising Scholars Conference 2024 and WEFI (Workshop for Entrepreneurial Finance and Innovation) Fellows 2023 for helpful comments.

1 Introduction

> *We are in the final 10 years of venture capital as we have come to know it. AI is going to remake the startup industrial complex, from its core. Venture firms will have to remake themselves into a combination of people and AI.* – James Currier General Partner at NFX

> *(AI) is moving from a nice-to-have to a must. And you need to be data-driven because everything is larger, faster, earlier; there are so many funds around and differentiation is hard to show sometimes, and finally because quality is increasing overall, and to win the best deals and allocate your capital wisely you can't simply trust your gut anymore.* – Francesco Corea Director of Data Science at Greycroft

The emergence of digital data and machine learning techniques has led to the increase in adoption of data technologies in financial markets (Heath (2019)). Prior literature and the popular press cite three main advantages of using data technologies. First, data technologies have proven successful for prediction problems based on common inputs. For example, data technologies have been useful in residential markets for identifying which houses are likely to appreciate (Raymond (2024)). Second, data technologies can help mitigate human biases in the investment decision process. Empirical evidence on this is mixed; some findings support this claim (D'Acunto, Ghosh, and Rossi (2022)), while other work highlights that training algorithms on historical data can perpetuate biases, widening the gap between various groups of people (Fuster et al. (2022)). Third, and less empirically explored, is the use of data technologies to scale an investors' investment opportunity set (Rasouli, Chiruvolu, and Risheh (2023)). This occurs during the deal sourcing and collection stage. Algorithms and automation processes can decrease time and increase efficiency, as traditional methods such as searching online, networks and marketing campaigns are limited in how they provide access to more investment opportunities (Majbour, Forbes (2023)).

This paper examines whether data technologies broaden investment opportunities for financial intermediaries, specifically in the context of the Venture Capital (VC) industry. The VC market provides an interesting setting for four reasons. First is the importance of the industry as VCs are crucial providers of capital to young, innovative firms with approximately 50% of all publicly traded companies having been VC-backed at some point prior to the IPO (Gornall and Strebulaev (2021)). Second, investing in young startups hinges on timely information for deal flow of new and unique businesses, an area where data technologies can provide significant advantages. VCs' deal sourcing processes have been subject to criticism, as they are mostly manual and inbound (i.e. from the VCs' professional network, Gompers et al. (2020)), leading to incomplete coverage of potential investment opportunities. Third, VCs are increasingly adopting statistical techniques (see Figure 1) with survey evidence suggesting that 75% of VCs will use data technologies in some capacity to influence investment decisions by 2025 (Gartner (2023)). Lastly, prior research on data technologies in the VC industry focuses on deal screening (Bonelli (2023)) and VC biases in the investment decision process (e.g. Lyonnet and Stern (2022)), however little is known how data technologies impacts deal sourcing, one of the most important steps in the VC investment decision process (Sørensen (2007), Gompers et al. (2020)). I therefore ask two research questions: (1) do data technologies broaden VCs' investment opportunity set and (2) does this have spillover effects on which startups receive funding in the VC industry.

To identify if and when VCs adopt data technologies, I utilize detailed employee data from Crunchbase and LinkedIn. The rationale is that VCs using data technologies rely on human capital and expertise to implement the data infrastructure. Prior research has used job postings to infer technology adoption in other settings (e.g. Alekseeva et al. (2021) and Goldfarb, Taska, and Teodoridis (2021)) and specifically in VC literature (Retterath (2020), Bonelli (2023)). Using job titles and descriptions from a complete history of VC employees, I identify when VCs hire data scientists and classify a VC firm as data-driven from the date of its first data-related employee hire[1].

In the first part of the paper, I investigate whether data technologies impact VCs' investment opportunity set. I use the geographic concentration of the VC industry in the US as my empirical setting. Specifically, the VC industry in the US is geographically concentrated (Chatterji, Glaeser, and Kerr (2014), Chattergoon and Kerr (2022)), with 79% of total venture capital invested startups located in California, Massachusetts, and New York (NVCA (2019),

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

Chen and Ewens (2021)). VCs tend to also locate in these clusters and thus invest locally as geographic boundaries facilitate information transmission amongst VC networks (Chen et al. (2010)), with the likelihood of investment decreasing in distance (Sorenson and Stuart (2001)). My overarching prediction is that after VCs adopt data technologies, they are no longer limited to innovation clusters to find investment opportunities as using data technologies allows them to find all potential investments with an online presence. In my first empirical test, I examine whether VCs become more likely to invest in startups located in areas with low VC activity after adopting data technologies. The intuition is that areas with low VC activity are less likely to have startups part of established VC networks (Hochberg, Ljungqvist, and Lu (2010)). I find that, after adopting data technologies, VCs are 9-13% more likely to invest in commuting zones in the lowest decile of historical VC activity and 10-30% more likely to invest in commuting zones with 25 or fewer VC investments in the previous 5 years. I find similar results when conducting the same analysis at the state-level. Relatedly, I find that after VCs adopt data technologies, they are 12% more likely to invest in distantly located startups (i.e. in the top tercile of distance). These findings provide evidence that data technologies increase VCs potential investment opportunity set to startups that would otherwise be excluded from their professional networks.

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

In sum, my findings indicate that VCs are more likely to invest in startups that would otherwise fall outside of their professional networks, providing evidence that data technologies increase investors opportunity sets for sourcing investments. I implement various strategies to mitigate concerns that results are driven by correlated unobservables. To address concerns that VCs that use data technologies are different from those who do not, I include VC firm fixed effects, therefore comparing VC investment decisions before and after technology adoption. In addition, I include industry × funding round stage × investment year fixed effects with time varying commuting zone level controls (or commuting zone × investment year fixed effects depending on the outcome variable) to alleviate concerns of the startup's local time trends coinciding with VC's adoption of data technologies that leads them to invest in more distant startups. Finally, results also hold after excluding investments after 2019. COVID-19 provided a shock to in-person interactions, and recent studies find that VCs tend to invest in more distantly located startups after March of 2020 (Han et al. (2022), Alekseeva et al. (2022)).

While the stringent specifications and robustness tests mentioned above address many endogeneity concerns, I conduct an additional analysis to further mitigate potential issues with omitted variables. Specifically, I isolate variation in VCs' data technology adoption that stems from early exposure to AI, removing potential bias from demand shocks driving firms' technology adoption and investment strategies. The intuition is that commercial interest in AI became widespread in the 2010s with technology firms first introducing AI into products for consumers (e.g. Apple introducing Siri in 2011) and later non-technology firms using AI to enhance business operations (e.g. Walmart using cameras on floor scrubbers to determine real-time inventory levels in 2017). However, with any new technology, there are some VC firms whose adoption costs are lower than others. I posit that VCs who invested in firms specializing in AI prior to 2010 would have early exposure to AI and thus an earlier understanding of its advantages. Since VC-backed startups pioneered many developments in AI, they would be some of the first financial intermediaries to be exposed to the technology[2].

[2] For example, in 2005, VCs invested in Predictix, which focused on offering clients big data and analytics processes to forecast future business operations. Similarly, in 2008, VCs invested in Voci which pioneered the speech-to-text algorithms in hardware.

Thus for each VC, I compute a measure of VC exposure to AI through their investments in startups in AI-related industries before 2010. I instrument VCs adoption of data technologies using this exposure measure. The instrument meets the necessary requirements for an IV approach, demonstrating a strong first stage (relevance condition) and only influencing VCs' deal-sourcing post-2010 through the VCs' adoption of data technologies (exclusion restriction). I show that the instrumented adoption of data technologies predicts an increase in investments in areas with a low VC activity after 2010. My findings provide added confidence that VCs' data technology adoption has causal effects on the types of investments they choose to make, further supporting the claim that date technologies increase a VCs' investment opportunity set.

In the second part of the paper, I investigate if data technology adoption impacts the overall geography of innovation. As previously mentioned, entrepreneurial and venture activity is highly concentrated in the US. While there are increasing returns to scale for entrepreneurial activity in innovation hubs, there is a growing concern that this concentration in activity can lead to the "hollowing out" of innovative activities in other parts of the country (Lerner and Nanda (2020), Glaeser and Hausman (2020)). I therefore examine whether there are persistent effects of entrepreneurial and venture activity in areas of low VC activity that data-driven VCs choose to invest in. I start by constructing a panel of commuting zones across my sample period that received 25 or less VC investments over the last 5 years. I then identify startups in these commuting zones that receive funding for the first time from data-driven VCs and classify these commuting zones as treated, a total of 26 commuting zones. Using a stacked difference-in-difference framework, I then examine whether commuting zones that receive investment from a data-driven VC are likely to experience an increase in VC activity in subsequent years compared to commuting zones that did not receive investments from data-driven VCs.

I begin with the startup's side. I find that after a data-driven VC invests in a commuting zone for the first time, the number of startups that receive their first VC financing increases by 22-29% in the next five years compared to commuting zones that did not receive financing from data-driven VCs. I also find that the number of patents filed by startups backed by VCs increases by 31% and while insignificant, I find that the number of patents filed by entrepreneurial startups increases by 4%. From the VC side, I find that after a data-driven VC invests in a commuting zone with low VC activity, the number of funding rounds in that commuting zone increases by 12-14%, the number of unique VCs investing in that commuting zone increases by 23-29% and the number of VCs investing for the first time in the commuting zone increases by 45-54%. I include commuting zone and year fixed effects as well as pre-strengthening commuting zone-level controls, income, GDP, and percentage of college graduates, to account for the possibility that commuting zones with certain characteristics experience a change in outcomes post data-driven entry. In additional analyses, results are robust to dropping control commuting zones that experience zero VC investments in the previous five years, mitigating concerns that results are driven by mechanical effects. Overall, these results suggest that entry of data-driven VCs has long-lasting effects on the entrepreneurial and venture activities in those commuting zones. This can have important policy implications for areas hoping to attract VC funding.

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 discusses the potential impact of data technology adoption on the geography of the VC industry. Section 7 concludes.

2 Contribution to Prior Literature

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

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

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

3 Institutional Background - Traditional VC Model vs Data Driven VCs

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

Specifically for sourcing deal flow, traditional VCs rely on their internal networks and reputation in localized markets to find investments. Nearly 50% of VC investments are syndicated (Lerner (2022)), indicating that a lead investor reached out to other VCs to invest a specific firm. While better networked VCs are shown to have superior fund performance (Hochberg, Ljungqvist, and Lu (2007)), the coverage of potential investments is largely incomplete which could prevent best possible fit between startups and VCs. However using data technologies allows the VC to find every company possible. VCs can use data technologies to identify firms at their earliest stage, through trending repositories on Github, webcrawlers finding new websites products launched on commercial databases, new LinkedIn profiles or new registrations of financing on public registers (Retterath, 2020b). Once the firms are identified, VCs can use data technologies to collect as much information on each firm to create a company profile by scraping company websites or LinkedIn and Twitter profiles or using APIs from commercial databases such as Crunchbase, Pitchbook, AngelList, and CB Insights to name a few[3].

[3] See https://medium.com/birds-view/the-future-of-vc-augmenting-humans-with-ai-30f1d79a09c3 for

Data technologies then use all the gathered information to score the startups and provide informative metrics for VCs to decide which companies to invest in. While this paper focuses mostly on sourcing startups, see Bonelli (2023) for more information on how data technologies are used in the screening portion of the pre-investment screening of startups.

4 Data and Summary Statistics

4.1 VC Investments

I use data from Crunchbase to construct my investment sample. Crunchbase is an online database providing detailed information on startup firms and their investors. I start by defining my VC investor sample. I keep all VC firms headquartered in the US and defined as venture capitalists, micro venture capitalists or private equity firms [4]. I then merge the remaining VCs with Preqin and VentureXpert to ensure coverage in multiple databases. I am left with 1,985 distinct VC firms during my sample period of 2000 to 2022 [5]. For each VC firm, I gather information on their founding year, headquarter location, 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 2000. I restrict my sample of investments to those classified as pre-seed, seed, and series a, b, c, and d+. I assume that VCs use data technologies to identify firms for first-time investments and exclude follow-on investments when testing the impact of data technology adoption on certain outcomes. My final sample amounts to 78,445 first time investments.

Lastly, I gather information on all the startups invested in by my VC sample. This includes their founding year, industry classification, head quarter location and founder information from their employee and job histories. Founder information includes gender, education, and whether they are a serial entrepreneur. My final sample includes 29,375 distinct startups that

more information on the sourcing and screening process of data-driven investments.

[4] 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.

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

were at some point VC funded.

4.2 Methodology to Identify Data-Driven VCs

Following prior literature (e.g. Bonelli (2023), Retterath (2020), Raymond (2024)), I identify VCs using data technologies as those who hire data scientists or data related employees. The rationale is that VCs using data technologies rely on human capital and expertise to implement the maintain the data infrastructure. Crunchbase includes data on employees at VC firms, including executives, partners, analysts, advisors, engineers, and other professional staff. For each VC firm in my sample, I collect the whole history of jobs at the firm in Crunchbase (including past jobs that are no longer active). For each job, Crunchbase reports the starting date, end date (or whether the job is current), the job type (employee, executive, advisor, board member or observer), the job title, and the unique identifier of the person. In addition, I collect each VCs' LinkedIn URL. I then create a sample of VC employees using data from Crunchbase and scraping LinkedIn job histories.

Next, I identify an initial list of data-driven VCs. Using my sample of VC employees, I clean all job titles and descriptions using standard text cleaning procedures. Then, I search for each word in my data-related job list in the full set of job titles in my VC sample, giving me a final list of VC firm job with data-related titles. I only keep job types listed as "employee" or "executive" to ensure my list does not capture advisors or board members. I also remove jobs associated with people who advise startups and lastly do a manual check to ensure each job is associated with the use of the technology in the pre-investment screening process[6]. Following prior literature, I use the starting date of a VCs first data-related employee hire to classify a VC as data-driven. Otherwise, VCs are considered traditional. I am able to identify 54 VCs that adopt data technology in my sample which corresponds to 3,273 data-driven investments.

[6] As a final sanity check, I compare my list of data-driven VCs with those identified in https://www.datadrivenvc.io/, a website created by Andre Retterath, a data-driven VC in Europe, of which I have considerable overlap.

5 Data-Driven Investors and Investment Opportunities

In this section, I investigate whether data technologies lead investors to fund startups typically outside of their professional networks, providing evidence that data-driven investing leads to an increase in VCs' investment opportunity set.

5.1 Areas with Low History of VC Activity

First, without claiming causality, I investigate whether data technologies scale investors' opportunities set. Without access to deal flow data, I proxy for this by identifying investment characteristics that would be outside a VCs' typical network. The intuition is that data technologies are able to find all possible startups with an online presence and can therefore identify potential investments that would fall outside a VCs network. Since the VC industry is highly concentrated with over 79% of capital invested in California, New York, and Massachusetts (Lerner (2010)), I assume that startups located in areas with a low history of VC activity to fall outside a VCs professional network.

I classify a startup as being outside a VCs network if they are located in a commuting zones (state) in the bottom decile of VC investments in the last 5 years. I then estimate the following regression at the investment level:

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

The dependent variable is an indicator if the startup is located in a commuting zone (state) in the bottom decile of VC investments in the previous 5 years. The main explanatory variable, $DataDriven$, is a dummy variable equal to 1 if VC j is classified as data-driven as of the investment date and 0 otherwise. $X_{j,k,t}$ are time varying controls of VC j and startup k. α_j are VC firm fixed effects and α_c are commuting zone fixed effects to control for any time invariant VC or commuting zone characteristics. $\gamma_{i\times s\times t}$ are startup industry $i \times$ funding stage $s \times$ funding year t fixed effects to alleviate concerns of the startup's time and industry trends coinciding with VC's adoption of data technologies that leads them to invest in startups located in low activity areas. The coefficient β is therefore estimated by comparing VC j's investments before versus after data technology adoption relative to other VC firms' investments in the same industry-stage-year segment. Standard errors are double clustered at the VC-firm year level.

Table 2 Panel A reports the results. Column (1) estimates Equation 1 without VC firm fixed effects. The coefficient on $DataDriven$ is positive and significant, indicating that data-driven VCs are 7% more likely to invest in startups located in low activity commuting zones. Column (2) includes VC firm fixed effects. The coefficient on $DataDriven$ can be interpreted as after VCs adopt data technologies, they are more likely to invest in startups located in low activity commuting zones. In Column (3), I include time varying commuting zone characteristics, the natural log of GDP and income ($Log(GDP)$ and $Log(Income)$ and the percentage of the population with a four-year college degree ($PercCollege$), to control for any local market trends that may attract VC attention. The result remains consistent and taken together can be interpreted as after VCs adopt data technologies, they increase their investments in startups located in low activity commuting zones by 9-13%. In columns (4) through (6) I repeat the analysis at the state level. Data-driven VCs are 11% more likely to invest in states with low activity (column (4)) and after VCs adopt data technologies, they are 13-25% more likely to invest in low activity states.

In addition, I estimate the following event study specification:

$$Y_{j,k,t} = \sum_{l=-4,l\neq-1}^{5+} \beta_l \{DataDriven(l) + X_{j,k,t} + \alpha_j + \alpha_c + \gamma_{i\times s\times t} + \epsilon_{j,k,t} \tag{2}$$

where $\{DataDriven(l)$ is a dummy variable equal to one if VC j adopts data technologies over the sample period and if year t corresponds to l years before/after the technology adoption. The omitted category is the year before the adoption. Panel A of Figure **??** illustrates the dynamic effects showing the estimated coefficients of β_l for startups located in commuting zones with low VC activity and Panel B in states with low VC activity. These figures complement the results that VCs invest in startups located in areas with little history of VC activity.

In Panel B of Table 2 I classify commuting zone (states) as areas with low VC activity if they received 25 or fewer investments over the past five years. I chose 25 as prior literature has classified formal VC markets at the state and MSA level as receiving more than 25 investments over a five year horizon (Hochberg, Ljungqvist, and Lu (2010)). The point estimate on $DataDriven$ in columns (2) and (3) of Panel B can be interpreted as after VCs adopt data technologies, they are 2.5-15% more likely to invest in low activity commuting zones. While insignificant, the point estimates on $DataDriven$ in columns (5) and (6) can be interpreted as after VCs adopt data technologies, they are 10-40% more likely to invest in low activity states. Taken together, these results provide evidence that VCs are more likely to invest in startups not found in their traditional networks, suggesting that data technologies scale their investment opportunity set.

5.2 Other Proxies for Out-of-Network Investments

In addition, I use other proxies for startups considered to be outside VC networks. VCs tend to locate in innovation clusters and invest locally as geographic boundaries facilitate information transmission within VC networks (Chen et al. (2010)) with the likelihood of investment decreasing in distance (Sorenson and Stuart (2001)). I therefore classify startups located faraway as those outside traditional VC networks. Specifically, I classify an investment as distantly located if it is in the top tercile of distance over my sample period. I estimate Equation 1 and column (1) of Table 3 displays the results. The coefficient is positive and statistically significant. In column (2), interact industry × stage × year fixed effects with the startup's commuting zone to control for time varying local shocks. Results are consistent and can be interpreted as, after VCs adopt data technologies they are 10% more likely to invest in distantly located startups.

In columns (3) and (4), I replace the outcome variable with an indicator if the VC invests in a different industry than their specialization. A large literature shows that VCs tend to specialize in investing in various industries (e.g. Hochberg, Mazzeo, and McDevitt (2015)) and these industries can form established networks within the VC industry (Hochberg, Ljungqvist, and Lu (2010)). I therefore classify a VC as specializing in a particular industry if more than 40% of their investments in the previous 5 years are in startups from the same industry. Crunchbase uses a granular industry specification system with over 750 industry classifications. Using a supervised machine learning approach, I classify these into 7 industry groups (Figure A1): Software and IT, Health Care and Biotechnology, Hardware and Electronics, Financial Services, Business Services, Consumers, Industrial and Energy. The coefficient on $DataDriven$ in columns (3) and (4) is positive and statistically significant, and can be interpreted as, after VCs adopt data technologies they are 7.5% more likely to invest in a startup in a different industry than their specialization, a 40% increase from the unconditional mean.

Lastly, in columns (5) and (6), I investigate whether VCs rely less on other investors to find startups and invest with. I therefore replace the outcome variable with an indicator equal to 1 if a VC invests with another VC located in the same state as the startup, conditional on investing out of their headquartered state. The intuition is that VCs who invest outside of their home state are less likely to know of potential investment opportunities in other VC markets unless they know another VC located close to the startup. However, if VCs use data technologies to find investments, they can now find the startup without local help. The coefficients on $DataDriven$ are negative and statistically significant and can be interpretted as, after VCs adopt data technologies, and conditional on investing outside of their headquartered state, they are 3-5% less likely to syndicate with a local VC, a 4-7% decrease from the unconditional mean.

5.3 The Causal Effects of Adopting Data Technologies

My results so far do not speak to whether there is a causal link between data technologies and expanding investment opportunities. While my time varying controls and fixed effect specifications control for any time trends coinciding with VC's adoption of data technologies in the startups locations, a major concern is that data technology adoption is an endogenous decision that may be correlated with unobserved changes at the VC firm, resulting in an omitted variable bias. One potential omitted variable is the impact of the COVID-19 pandemic on VC investment decisions. COVID-19 occurred at a similar time to a large shift in artificial intelligence investments as well as limited human interaction between VCs and startups. Recent research by Alekseeva et al. (2022) and Han et al. (2022) find that after the pandemic, VCs invested more in distantly located startups. My results hold after excluding investments made after 2019 in my baseline specification (see Table A1 and Table A2 for reference). In the event of other omitted variables, I develop an empirical strategy to estimate the causal impact of adopting data technologies on VC investments. My approach is to isolate variation in VCs' data technology adoption that comes from early exposure to AI, mitigating potential bias from demand shocks driving firms' technology adoption and investment strategies.

5.3.1 Identification Strategy

Commercial interest in AI became widespread only around 2010 with technology firms first introducing AI into products for consumer (e.g. Apple introducing Siri in 2011) and later non-technology firms using AI to enhance business operations (e.g. Walmart using cameras on floor scrubbers to determine real-time inventory levels in 2017). In 2012, researchers from Google introduced a deep Cognitive Computation Neuroscience (or CNN) architecture that won the ImageNet challenge and triggered the explosion of deep learning research and implementation (Krizhevsky, Sutskever, and Hinton (2012)). Many firms - both in and out of the technology sector - have adopted these methodological advances in their business operations since then. Recent research by Babina et al. (2024) finds a large increase in AI investments by public firms across industries, leading to growth in sales, employment, and market valuations. For early adopters of AI in industries other than technology, such as VCs in the financial services industry, firms to understand the benefits of implementing this technology as well as the know how to do so. While AI became popular for commercial use after 2010, young, innovative startups were some of the first firms to pioneer AI's development in the 2000s (for example, Predictix founded in 2005 offers clients big data and analytics processes to forecast future business operations and Voci founded in 2008 that pioneered the speech-to-text algorithms in hardware). VCs that invested in these startups would have a first movers advantage in terms of understanding the uses of AI ahead of other VCs and investors. I hypothesize that VCs who invested in startups specializing in AI prior to 2010 are likely to be early adopters of data technologies and change their investments inline with my previous findings.

AI Industry Exposure I exploit the cross-sectional heterogeneity in the impact of AI industries to identify the effect on data technology adoption by VCs. Crunchbase categorizes companies into 750 industries to account for heterogeneity across startup's specific market segments[7]. Following methodology used in Bonelli (2023), I assign a treatment intensity to each industry in Crunchbase proxying for the extent to which that industry would specialize in artificial intelligence. To create industry-level treatment intensities, I rely on business descriptions of firms in the Crunchbase database, including those of firms that were not VC-funded (Crunchbase covers other types of firms - including public and private that are \were not necessarily VC-backed). I start by collecting AI terms defined in the Artificial Intelligence Glossary from Tech Target, a marketing company that provides data-driven services to business-to-business technology vendors[8]. Table A3 reports the terms contained in the glossary. They include keywords such as "Artificial Intelligence", "Machine Learning" and "Natural Language Processing". I then search for these terms in the business descriptions of all companies in Crunchbase[9]. Finally, for each industry I compute the fraction of company descriptions featuring at least one AI term and I rank industries according to this metric. I only consider industries with more than 100 business descriptions to avoid assigning industry-level treatment intensities that are too dependent on a few companies. Treatment intensity (between 0 and 1) is then defined as the overall percentile rank in the industry distribution:

$$IndustryExposure_i = \text{Rank}_I\{\frac{\text{Nb. Company Descriptions with Match in Industry } i}{\text{Nb. Company Descriptions in Industry } i}\} \tag{3}$$

where I is the set of industries in Crunchbase. Intuitively, industries in which companies mention AI terms more often are more likely to be part of the AI industry. Panel A of Table A4 shows the ten industries with the highest treatment intensities. It includes industries such as "Machine Learning", "Artificial Intelligence", "Natural Language Processing", and "Text Analytics". The least exposed industries are presented in Panel B and encompass industries such as "Timber", "Bakery", and "Laundry". This is not surprising as companies in these industries

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

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

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

are less likely to benefit from AI.

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

$$VCExposure_j = \frac{1}{N_{j,2010}} \sum_{i \in A_{j,2010}} IndustryExposure_i, \tag{4}$$

where J is the set of VCs with investments before 2010, $A_{j,2010}$ is the set of investments made by VC firm j before 2010, $N_j 2010$ is the number of investments in this set, $IndustryExposure_i$ is the treatment intensity of the industry of the startup corresponding to investment i, defined in Equation 3. VC firms with the highest exposure are those with most of their investments before 2010 in industries with high treatment intensity, creating within-industry variations across investments made by investors with different VC-level exposures.

5.3.2 Instrumental Variables Approach

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

$$DataDriven_{j,k,t} = \beta\{VCExposure_j \times Post_t\} + X_{j,k,t} + \alpha_j + \alpha_c + \gamma_{i\times s\times t} + \epsilon_{j,k,t}, \quad (5)$$

where $DataDriven_{j,k,t}$ is an indicator if the investment was made by a data-driven VC j in startup k in year t $VCExposure_j$ is a VCs' exposure to AI through their investment prior to 2010 as defined in Equation 4. $Post_t$ is a dummy equal to one after 2010 and zero otherwise. $X_{j,i,t}$ are time varying controls VC and startup controls. α_j are VC firm fixed effects. α_c are startup commuting-zone fixed effects. $\gamma_{i\times s\times t}$ are startup's industry × commuting zone × funding year fixed effects. Standard errors are clustered at the VC firm and year level.

The results of the first stage are displayed in column (1) of Equation 4. The coefficient on $VCExposure_j \times Post_t$ is positive and statistically significant and the F-statistic 12.45, greater than the conventional level of 10. In addition, I performed the following event-study difference-in-difference specification.

$$DataDriven_{j,k,t} = \sum_{l=-4,l\neq-1}^{5+} \beta_l\{VCExposure_j \times Year(l)_t\} + X_{j,k,t} + \alpha_j + \alpha_c + \gamma_{i\times s\times t} + \epsilon_{j,k,t}, \quad (6)$$

where $Year(l)_t$ is a dummy variable equal to one if year t corresponds to l years before/after 2010. The omitted category is year 2009. Figure 3 graphs the estimated β_l in equation Equation 6. It shows no pre-trend. The increase in the likelihood of observing an investment made by a data-driven VC shows up in the years after 2010 and persists even 10 years after. Taken together with the strong first-stage, this satisfies the relevance condition for the instrument.

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

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

where $Data\hat{D}riven$ is instrumented by VCs' exposure to AI prior to 2010 and $Y_{j,k,t}$ is an indicator equal to 1 if the investment is made in a startup located in a commuting zone or state with low VC activity. The empirical specifications in Equation 5 and Equation 7 require observing the industry composition of VCs portfolios before 2010. This analysis therefore consists of 44,683 first time VC investments by 739 VC firms. The summary statistics for this sample can be found in Panel B of Table 1. Columns (2) and (4) in Equation 4 show the OLS results using this sample. The results are similar to that of the baseline specification in Equation 2.

The results for the second stage can be found in column (3) for the commuting zone level and column (5) for the state level. The coefficients are positive and statistically significant, indicating that adopting data technologies does expand VCs' opportunity set as proxied by increased investment in areas with low VC activity.

Other Instrumented Results I repeat the above analysis with other proxies for investments outside of VC networks. The results are displayed in Equation 5. Column (1) shows the first stage. The coefficient on $VCExposure_j \times Post_t$ is positive and statistically significant and the F-statistic 13.68, supporting the relevance condition. The instrumented results for distantly located startups, investing in a different industry, and investing with a local VC syndicate are shown in columns (3), (5), and (7) respectively. The results indicate that data technology adoption results in VCs investing in startups located further away, in industries other than their specialization, and without local VCs, providing further evidence that data technologies expand VCs' opportunity set.

6 Implications for Areas with Low History of VC Activity

In the previous section, I demonstrated that data technologies expand VCs opportunity set as proxied by their investments in startups outside of their typical VC networks. In the main specification, I find that VCs are more likely to invest in startups located in areas with little history of VC activity. Since VC activity is largely concentrated in US, there is a growing concern that this can lead to the "hollowing out" of innovative activities in other parts of the country Lerner and Nanda (2020), Glaeser and Hausman (2020)). If data technologies are able to identify startups in need of funding anywhere in the country, this could prove a useful tool to extend entrepreneurial funding to areas outside the major VC hubs. Thus, in the second half of the paper, I examine the economic implications of data-driven VCs investing in low activity areas. I first describe the research design and then present my main findings.

6.1 Research Design

I begin by constructing a panel of all commuting zones in the US during my sample period that received 25 or fewer VC investments over the previous five years. I classify these commuting zones as areas with low VC activity as prior research establishes geographical markets to consist of more than 25 VC investments over a five year period (e.g. Hochberg, Ljungqvist, and Lu (2010)). I then identify startups in these commuting zones that receive funding for the first time by data-driven VCs and classify these commuting zones as treated, a total of 26 commuting zones. I classify all other commuting zones as my control group. I then construct a stacked difference-in-difference model, comparing various measures of VC activity before and after an investment made in the commuting zone by a data-driven VC. Specifically, I construct the following difference-in-difference regression:

$$Y_{d,c,t} = \beta\{Treated_{d,c} \times Post_{d,t}\} + \alpha_{d,c} + \alpha_{d,t} + \epsilon_{d,c,t}, \tag{8}$$

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

To ensure my specification satisfies the parallel trends assumption, I conduct the following specification:

$$Y_{d,c,t} = \sum_{l=-4, l\neq -1}^{5+} \beta_{d,l}\{Treated_{d,c} \times Year(l)_{d,t}\} + X_{j,k,t} + \alpha_{d,c} + \alpha_{d,t} + \epsilon_{d,c,t}, \qquad (9)$$

where $Year(l)_{d,t}$ is a dummy variable equal to one if year t in cohort d corresponds to t years before/after a commuting zone receives a data-driven investment.

6.2 Data-Driven Investment Entry and Entrepreneurial Activity

I start by looking at the impact of an investment by a data-driven VC in a commuting zone with low VC activity on entrepreneurial activity in subsequent years. Specifically, I look at the number of startups that receive their first ever VC financing, the number of patents filed by startups backed by VCs, and the number of patents filed by entrepreneurial firms. I following methodology introduced by Ewens and Marx (2024) to classify these patents as being filed by VC-backed startups or entrepreneurial firms. I run the specification outlined in Equation 8. The results are displayed in Panel A of Table 6. Column (1) and (2) show the results for the number of startups receiving their first VC financing. The coefficients is positive and statistically significant and can be interpreted as after a data-driven VC invests in a low activity commuting zone, that commuting zone experiences an increase of 22-29% of startups that receive their first VC financing compared to commuting zones that do not receive data-driven VC investment. In columns (3) and (4), the number of patents produced by VC-backed startups increases by approximately 31% and, while not statistically significant, the number of entrepreneurial firm patents increases by 4-5% in commuting zones that receive investment from a data-driven VC. The dynamics for Equation 9 are displayed in Figure 4, supporting the above results.

To mitigate concerns that counties that receive data-driven investments are different from control counties in the sense that while small, may have had a few VC investments prior to entry by a data-driven VC, I repeat the analysis but drop controls that never receive a a VC investment over the 5 years prior to data-driven entry. The results are displayed in Panel B of Table 6. The coefficient magnitudes and significance is similar to that of Panel A, and thus altogether, these results indicate that entry by data-driven VCs has a positve impact on future innovation output in areas with low VC activity.

6.3 Data-Driven Investment Entry and Venture Activity

Lastly, I look at the long-lasting impact of data-driven entry on low activity commuting zones that receive an investment by a data-driven VC compared to those who do not. Specifically I look at the number of funding rounds, the number of unique investors, and the number of unique investors that invest in the commuting zone for the first time after data-driven entry. I run the specification outlined in Equation 8. The results are displayed in Panel A of Table 7. In columns (1) and (2), I find that the number of funding rounds increases by 12-14% in commuting zones that experience an investment by a data-driven VC compared to commuting zones that do not. In columns (3) and (4), the number of unique investors investing in startups in treated commuting zones increases by 23-29% and columns (4) and (5), the number of unique investors investing in startups for the first time increases by 45-54% in treated commuting zones. The dynamics for Equation 9 are displayed in Figure 5, supporting the above results.

Similar to the entrepreneurial activity results, I mitigate concerns that counties that receive data-driven investments are different from control counties by dropping controls that never receive a a VC investment over the 5 years prior to data-driven entry. The results are displayed in Panel B of Table 7. The coefficient magnitudes and significance is similar to that of Panel A, and thus altogether, these results indicate that entry by data-driven VCs has a positive impact on future VC activity in a areas with low VC activity.

Overall, the increase in entrepreneurial activity and the increase in VC activity in low activity commuting zones after entry by a data-driven investor indicates that data technologies can have a positive impact on the financing of innovation in areas outside major clusters in the US.

7 Conclusion

The adoption of data technologies by VCs firms has the potential to significantly transform their investment strategies and the broader landscape of innovation. This paper demonstrates that data technologies enable VCs to broaden their investment opportunity sets, allowing them to identify and invest in startups beyond their traditional networks and geographic constraints. By leveraging detailed employee data from Crunchbase and LinkedIn, I track the adoption of data technologies and show that VCs become more likely to invest in areas with historically low VC activity, in distant locations, and in industries outside their previous specializations. These findings suggest that data-driven approaches can mitigate information frictions and enhance the efficiency of deal sourcing.

Further, the research indicates that the impact of data technologies extends beyond the immediate investment decisions of VCs. The entry of data-driven VCs into new geographic areas increases entrepreneurial activity and attracts additional VC investments, suggesting a potential deconcentration of innovation from traditional hubs to more diverse locations. This shift could have significant policy implications, highlighting the importance of supporting data technology adoption to foster a more equitable distribution of venture capital and innovation opportunities across different regions.

In conclusion, the adoption of data technologies by VCs not only enhances their ability to discover and invest in startups outside of their typical networks but also contributes to reshaping the geography of innovation. 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 vital. Future research should continue to explore the long-term effects of this technological shift on financial markets.

References

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

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

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

Alekseeva, Liudmila, Silvia Dalla Fontana, Caroline Genc, and Hedieh Rashidi Ranjbar. 2022. “From in-person to online: the new shape of the VC industry.” *Available at SSRN.*

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.

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

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

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

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

Chattergoon, Brad, and William R Kerr. 2022. “Winner takes all? Tech clusters, population centers, and the spatial transformation of US invention.” *Research Policy* 51 (2): 104418.

Chatterji, Aaron, Edward Glaeser, and William Kerr. 2014. "Clusters of entrepreneurship and innovation." *Innovation policy and the economy* 14 (1): 129–166.

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.

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

D'Acunto, Francesco, Pulak Ghosh, and Alberto G Rossi. 2022. "How costly are cultural biases? evidence from fintech."

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

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

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

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.

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.

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.

Garfinkel, Jon A, Erik J Mayer, Ilya A Strebulaev, and Emmanuel Yimfor. 2021. "Alumni Networks in Venture Capital Financing." *SMU Cox School of Business Research Paper,* nos. 21-17.

Glaeser, Edward L, and Naomi Hausman. 2020. "The spatial mismatch between innovation and joblessness." *Innovation Policy and the Economy* 20 (1): 233–299.

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

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

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

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

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. 2022. “The Death of Distance? COVID-19 Lockdowns and Venture Capital Investment.” *COVID-19 Lockdowns and Venture Capital Investment (July 13, 2022).*

Heath, Donald R. 2019. *Prediction machines: the simple economics of artificial intelligence: by Ajay Agrawal, Joshua Gans and Avi Goldfarb, Published in 2018 by Harvard Business Review Press, 272 pp., 30.00(hardcover),KindleEdition: 16.19, ISBN: 978-1-633695672.*

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

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

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

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

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

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

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

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

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

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

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

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

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

Rasouli, Mohammad, Ravi Chiruvolu, and Ali Risheh. 2023. "AI for Investment: A Platform Disruption." *arXiv preprint arXiv:2311.06251.*

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).*

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.

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 plots the number of data-driven investments over time (bars) and the percentage of total investments over time (line).

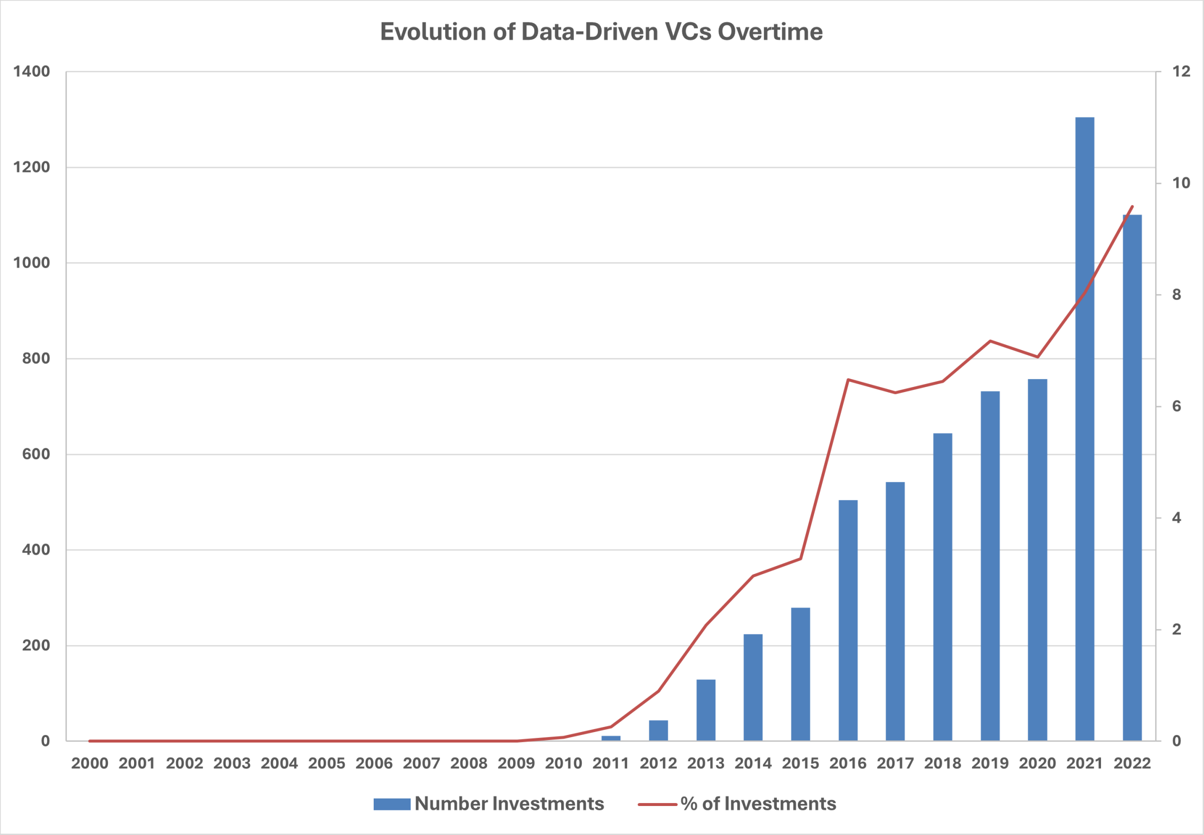
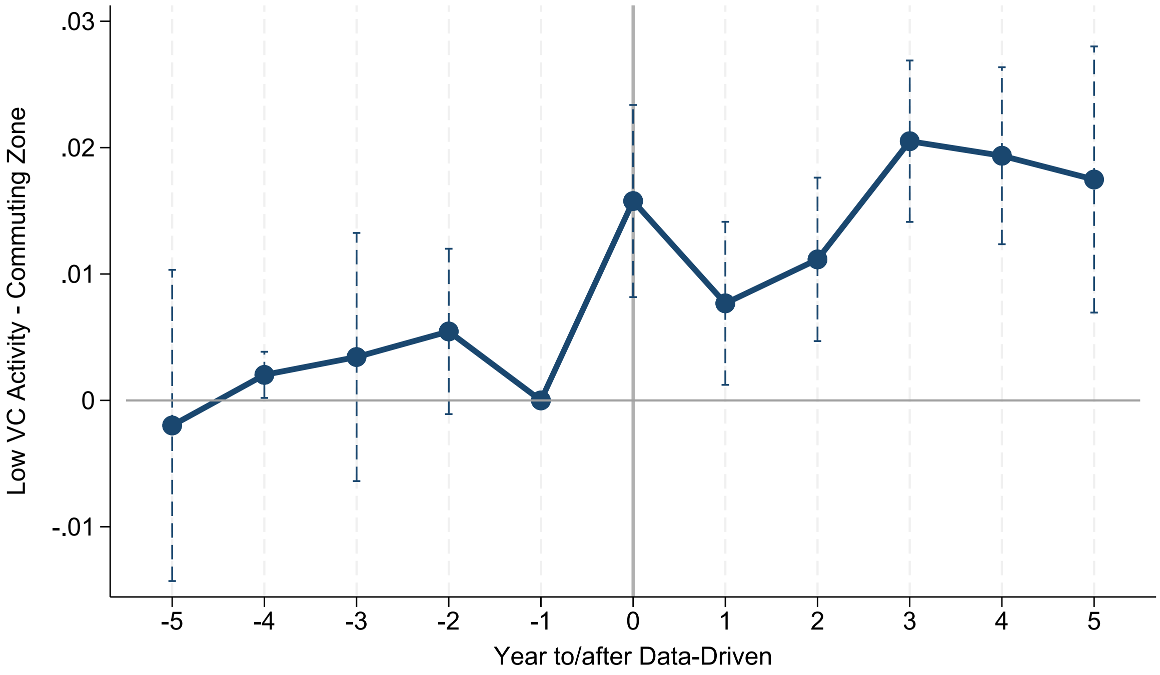


Figure 2: **Data Technologies with Areas of Low VC Activity**

The figures plot the estimated coefficients from Equation 2 at the VC-investment level, of each year relative a VCs' adoption of data technologies. In Panel A, the dependent variable is an indicator if the investment was made in a startup located in a commuting zone in the lowest decile of VC activity over the previous five years. In Panel B, the dependent variable is an indicator if the investment was made in a startup located in a state in the lowest decile of VC activity over the previous five years. The year prior to VCs adopting data technologies is the excluded category, reported as zero in the figures. The horizontal bars represent the 90% confidence interval for the coefficient estimates with standard errors double clustered at the VC firm-year level. Regressions include VC firm fixed effects, startup county fixed effects, and startup's industry-stage-funding year fixed effects. Regressions also control for the logarithm of the age of the VC firm and the logarithm of the startup's age. All control variables are measured at the time the investment is made.

(A) Commuting Zone



(B) State

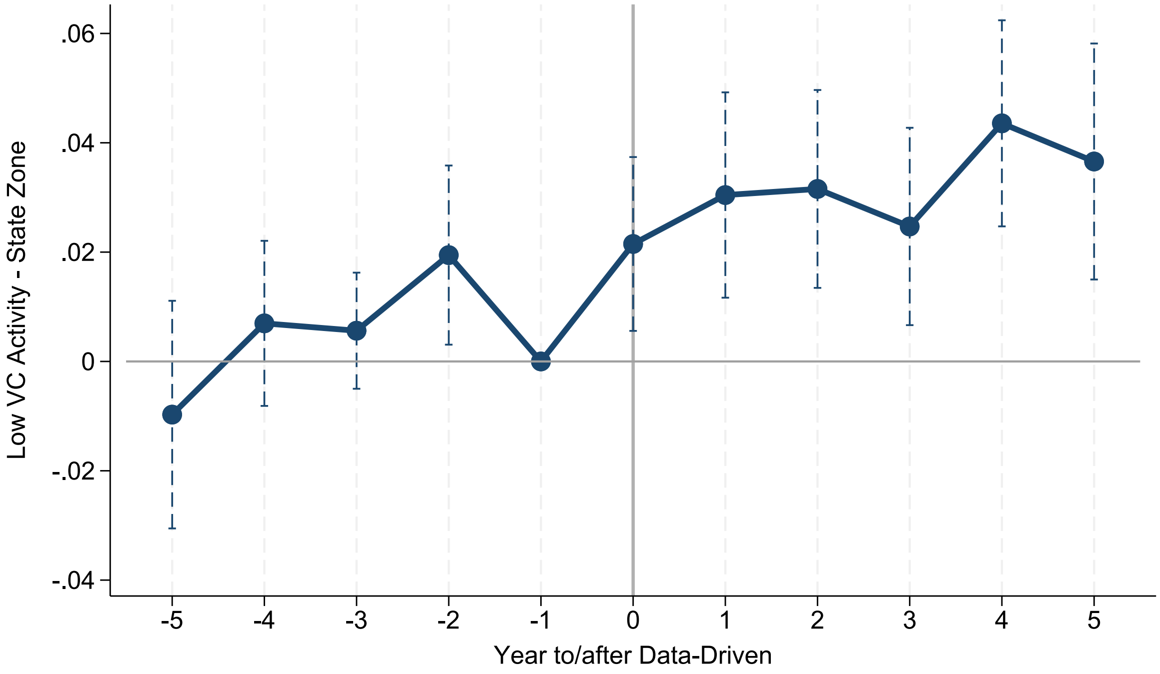


Figure 3: **Effects of VCs Pre-Exposure to AI Prior to 2010**

The figure plots the estimated coefficients from difference-in-differences regressions at the VC-investment level, for the interaction terms of each year relative to 2010 and the VC access to AI prior to 2010 (Equation 6). The dependent variable is a dummy indicating whether the investment is made by a VC classified as data-driven as of the investment date. The 2009 interaction term is the excluded category, reported as zero in the figures. The horizontal bars represent the 90% confidence interval for the coefficient estimates with standard errors double clustered at the VC firm-year level. Regressions include VC firm fixed effects, startup county fixed effects, and startup's industry-stage-funding year fixed effects. Regressions also control for the logarithm of the age of the VC firm and the logarithm of the startup's age. All control variables are measured at the time the investment is made.

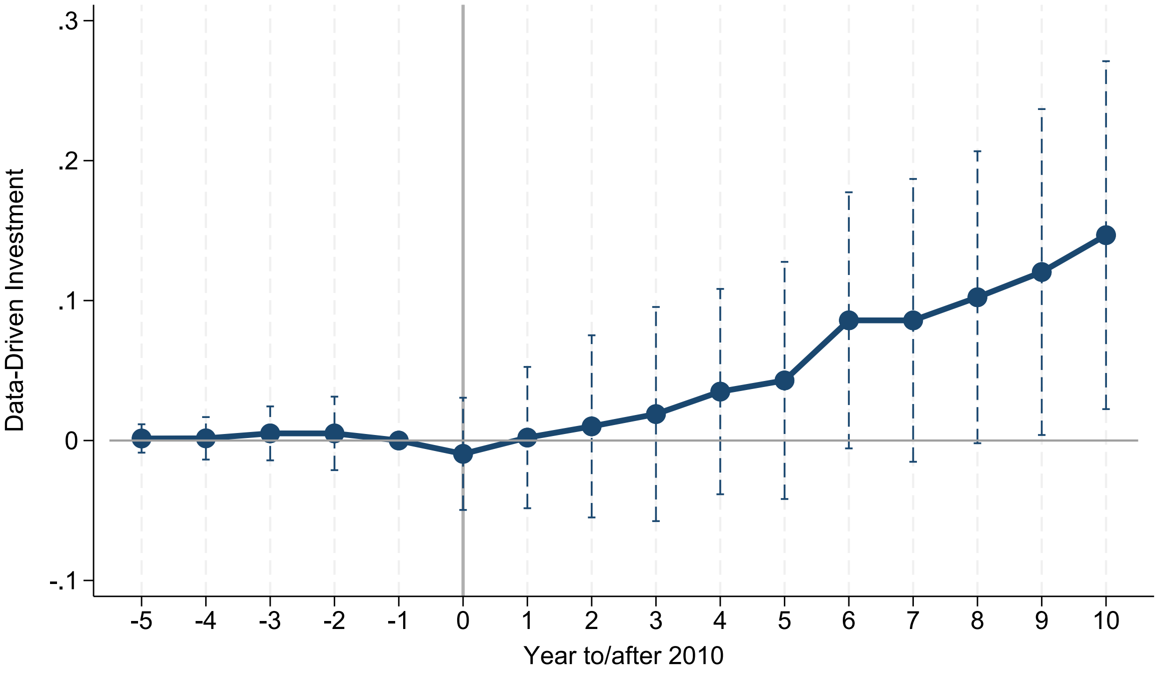
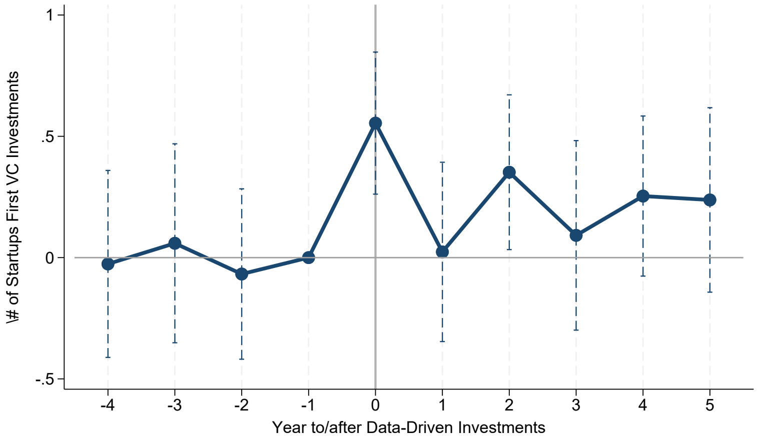


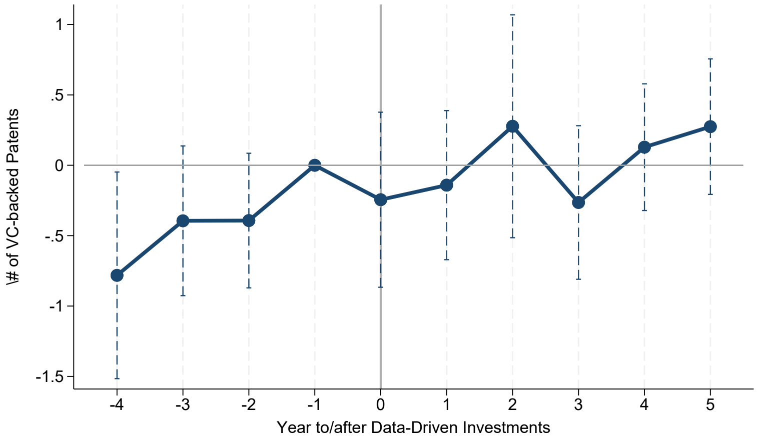
Figure 4: **Data Driven Investment Entry and Entrepreneurial Activity**

The figures plot the estimated coefficients from difference-in-differences regressions at the commuting-zone level, for the interaction terms of each year relative to data-driven entry into a low activity commuting zone (Equation 9). In Panel (A), the dependent variable is the number of startups receiving their first VC financing in the commuting zone. In Panel (B), the dependent variable is the number of patents filed by VC-backed startups. In Panel (C), the dependent variable is the number patents filed by entrepreneurial firms in a commuting zone. The year prior to data-driven entry is the excluded category, reported as zero in the figures. The horizontal bars represent the 90% confidence interval for the coefficient estimates with standard errors clustered at the county level. Regressions include cohort × county fixed effects and cohort × year fixed effects. Regressions also control for pre-data-driven entry VC financing.

(A) # First VC Financing



(B) # VC Patents



(C) # Entrepreneurial Patents

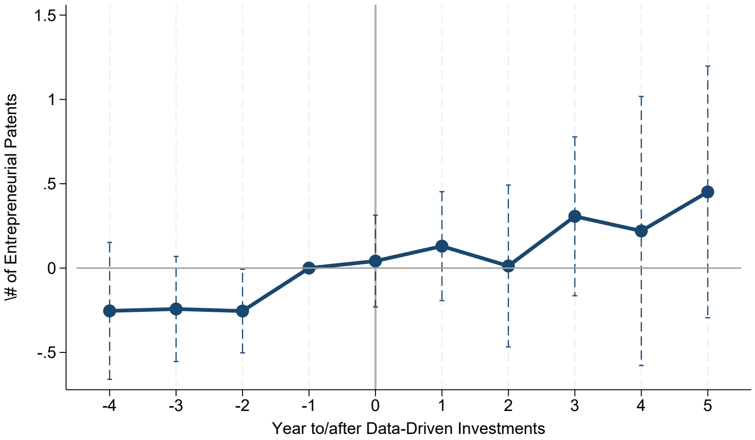
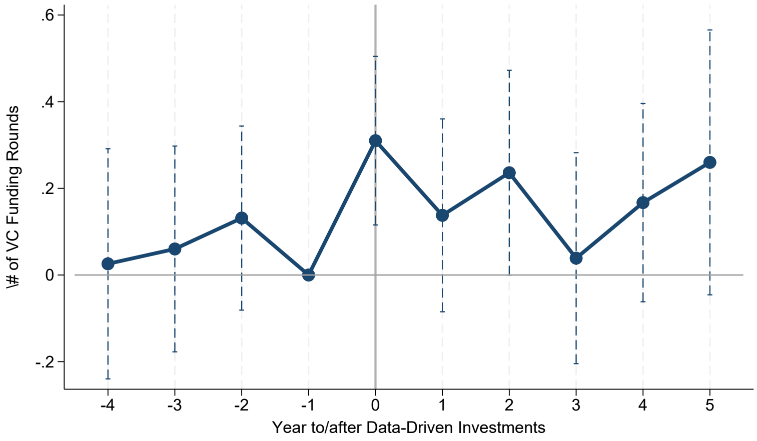


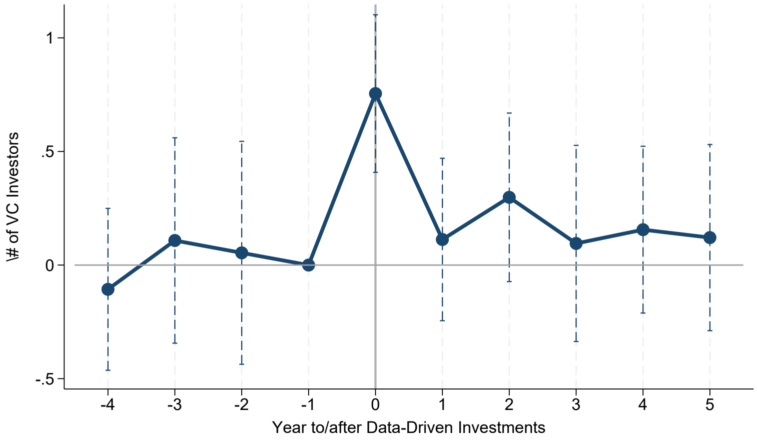
Figure 5: **Data Driven Investment Entry and VC Activity**

The figures plot the estimated coefficients from difference-in-differences regressions at the commuting-zone level, for the interaction terms of each year relative to data-driven entry into a low activity commuting zone (Equation 9). In Panel (A), the dependent variable is the number of funding rounds in the commuting zone. In Panel (B), the dependent variable is the number of unique VCs investing in the commuting zone. In Panel (C), the dependent variable is the number of unique first time investors investing in a commuting zone. The year prior to data-driven entry is the excluded category, reported as zero in the figures. The horizontal bars represent the 90% confidence interval for the coefficient estimates with standard errors clustered at the county level. Regressions include cohort × county fixed effects and cohort × year fixed effects. Regressions also control for pre-data-driven entry VC financing.

(A) # Funding Rounds



(B) # VC Investors



(C) # First Time VC Investors

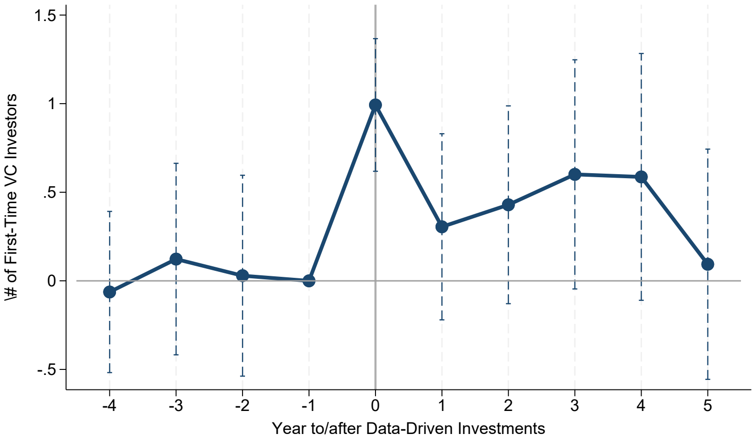


Table 1: **Summary Statistics**

	Mean	St. Dev.	P1	P25	Median	P75	P99	N
Panel A: First-Time Investment Level								
Data-Driven	0.04	0.20	0.00	0.00	0.00	0.00	0.00	78445
Startup Age	3.15	4.09	0.00	1.00	2.00	4.00	7.00	78445
Investor Age	12.53	15.07	0.00	3.00	7.00	16.00	30.00	78445
$\sum_{t=-5}^{-1}$# Funding Rounds (Commuting Zone)	1684	1523	4	355	1408	2565	4217	78445
$\sum_{t=-5}^{-1}$# Funding Rounds (State)	3637	3206	23	548	2792	6706	8577	78445
$\mathbb{1}(\leq 25$ Investments) (Commuting Zone)	0.04	0.20	0.00	0.00	0.00	0.00	0.00	78445
$\mathbb{1}(\leq 25$ Investments) (State)	0.01	0.11	0.00	0.00	0.00	0.00	0.00	78445
Distance (miles)	882	1031	0.00	16	322	1929	2567	78445
$\mathbb{1}$(Local Syndicate)	0.70	0.46	0.00	0.00	1.00	1.00	1.00	78445
$\mathbb{1}$(Different Industry)	0.17	0.38	0.00	0.00	0.00	0.00	1.00	78445
GDP millions (Commuting Zone)	77.76	56.02	3.73	39.69	62.75	95.60	176.35	78445
Income (Commuting Zone)	61,559	17,095	28,294	47,870	61,071	73,626	89,659	78445
Percentage College (Commuting Zone)	0.37	0.07	0.20	0.31	0.39	0.45	0.45	78445
Panel B: IV Sample								
Data-Driven	0.05	0.22	0.00	0.00	0.00	0.00	0.00	45125
VC Exposure	0.56	0.13	0.00	0.53	0.60	0.63	0.65	45125
Startup Age	3.15	4.50	0.00	1.00	2.00	4.00	7.00	44827
Investor Age	18.20	16.56	0.00	7.00	14.00	24.00	37.00	45125
$\sum_{t=-5}^{-1}$# Funding Rounds (Commuting Zone)	1432	1367	3.00	276	1138	2085	3584	45098
sum$_{t=-5}^{-1}$# Funding Rounds (State)	3167	2925	18	494	2496	5125	8333	45124
Distance (miles)	848.66	1026.17	0.00	16.83	288.76	1837.45	2567.23	44236
Local Syndicate	0.71	0.45	0.00	0.00	1.00	1.00	1.00	45125
Investment Outside Industry Specialization	0.19	0.39	0.00	0.00	0.00	0.00	1.00	45125
GDP mil (Commuting Zone)	69.62	51.60	3.62	39.40	54.42	83.30	155.86	43386
Income (Commuting Zone)	57,335	16,471	27,275	44,390	56,555	67,349	80,105	43386
Percentage College (Commuting Zone)	0.36	0.07	0.19	0.31	0.39	0.40	0.45	43356
Panel C: Commuting Zone-Level Sample								
# First VC Financing	0.28	0.79	0.00	0.00	0.00	0.00	1.00	47524
# VC Patents	0.16	0.88	0.00	0.00	0.00	0.00	0.00	44581
# Entrepreneurial Patents	1.16	6.60	0.00	0.00	0.00	1.00	3.00	44581
# Funding Rounds	0.91	2.17	0.00	0.00	0.00	1.00	3.00	47524
# Investors	0.60	1.78	0.00	0.00	0.00	0.00	2.00	47524
# Investors First	0.52	1.89	0.00	0.00	0.00	0.00	1.00	47524
$\sum_{t=-5}^{-1}$# Funding Rounds	0.71	1.67	0.00	0.00	0.00	1.00	2.00	47524
GDP mil	1.81	2.46	0.09	0.52	1.00	2.22	4.18	47524
Income	40,436	8,975	27,054	34,437	38,947	44,544	50,648	47,524
Percentage College	0.20	0.06	0.11	0.16	0.19	0.23	0.26	46764

Table 2: **Data Technology and Areas with Low History of VC Activity**

This table reports results for regressions at the VC-investment level, investigating whether the investments made by data-driven VCs after they adopt data technologies lead to different outcomes than those made by other VCs. In Panel A, columns (1) through (3), the dependent variable is an indicator if a VC made an investment in a startup located in a commuting zone in the lowest decile of VC activity. In Panel A, columns (4) through (6), the dependent variable is an indicator if a VC made an investment in a startup located in a state in the lowest decile of VC activity. In Panel B, columns (1) through (3), the dependent variable is an indicator if a VC made an investment in a startup located in a commuting zone that received 25 or fewer investments in the previous five years. In Panel B, columns (4) through (6), the dependent variable is an indicator if a VC made an investment in a startup located in a state that received 25 or fewer investments in the previous five years. All columns include startup commuting zone fixed effects and startup industry by funding stage by funding year fixed effects. Columns (2), (3), (5), and (6) include VC firm fixed effects. Control variables across all specifications include the logarithm of the age of the VC firm and the logarithm of the startup's age. Control variables in columns (3) and (6) include the natural log of startup county's GDP and income and the percentage of the population that received a college degree a year prior to the investment. Regressions are double clustered at the VC firm year level.

Outcomes:	$\mathbb{1}$(Low VC Activity CZ)			$\mathbb{1}$(Low VC Activity State)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Lowest Decile of VC Activity						
Data-Driven	0.007**	0.013***	0.009**	0.011**	0.025***	0.013**
	(2.42)	(4.77)	(2.79)	(2.67)	(3.61)	(2.78)
Log(Startup Age)	0.000	0.000	0.000	-0.001	0.000	0.001
	(0.10)	(0.33)	(0.23)	(-0.53)	(0.04)	(0.47)
Log(VC Firm Age)	-0.002*	-0.006	-0.002	-0.001	-0.004	-0.001
	(-2.05)	(-1.71)	(-1.01)	(-1.46)	(-0.93)	(-0.38)
Log(GDP)			0.336***			0.306***
			(5.52)			(3.12)
Log(Income)			-0.131			0.452
			(-0.93)			(1.65)
Perc College			-0.771*			0.950
			(-2.05)			(1.66)
VC-Firm FE	No	Yes	Yes	No	Yes	Yes
Org-Comzone FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.81	0.82	0.85	0.69	0.70	0.72
N	78445	78440	76485	78445	78440	76485
Panel B: 25 or Fewer VC Investments						
Data-Driven	0.004*	0.006*	0.001	0.005***	0.004	0.001
	(1.79)	(2.00)	(0.59)	(3.14)	(1.69)	(1.08)
Log(Startup Age)	-0.000	0.000	0.001	-0.000	-0.000	-0.000
	(-0.34)	(0.03)	(0.67)	(-0.75)	(-0.34)	(-0.76)
Log(VC Firm Age)	-0.001*	-0.004*	-0.002	-0.001	-0.003	-0.001
	(-1.91)	(-1.82)	(-1.06)	(-1.64)	(-1.34)	(-0.63)
Log(GDP)			0.309***			0.117***
			(7.09)			(7.11)
Log(Income)			-0.249***			0.016
			(-3.20)			(0.39)
Perc College			2.049***			1.001***
			(7.14)			(6.43)
VC-Firm FE	No	Yes	Yes	No	Yes	Yes
Org-Comzone FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.74	0.74	0.80	0.45	0.46	0.55
N	78445	78440	76485	78445	78440	76485

Table 3: **Data Technology and Other Measures of Out-of-Network Investments**

This table reports results for regressions at the VC-investment level, investigating whether the investments made by data-driven VCs after they adopt data technologies lead to different outcomes than those made by other VCs. In columns (1) through (2), the dependent variable is an indicator if a VC made an investment in a startup located in the top tercile of distance from their headquarters. In columns (3) and (4), the dependent variable is an indicator if a VC invested in a startup in a different industry from their specialization. In columns (5) and (6), the dependent variable is an indicator if a VC invested in a startup without a local VC syndicate, conditional on investing out of their headquartered state. All columns include VC firm fixed effects. Odd columns include startup commuting zone fixed effects and startup industry by funding stage by funding year fixed effects. Even columns include startup commuting zone by startup industry by funding stage by funding year fixed effects. Control variables include the logarithm of the age of the VC firm and the logarithm of the startup's age. Regressions are double clustered at the VC firm year level.

Outcomes:	$\mathbb{1}$(Top Tercile Distance)		$\mathbb{1}$(Diff Industry)		$\mathbb{1}$(Local Syndicate)	
	(1)	(2)	(3)	(4)	(5)	(6)
Data-Driven	0.040***	0.037**	0.073	0.081**	-0.034	-0.047**
	(2.92)	(2.68)	(1.61)	(2.08)	(-1.59)	(-2.19)
Log(Startup Age)	0.007	0.006	-0.003	-0.003	-0.051***	-0.059***
	(1.46)	(1.57)	(-0.95)	(-0.91)	(-11.99)	(-10.85)
Log(VC Firm Age)	0.024***	0.024***	0.065***	0.062***	0.022**	0.022**
	(2.89)	(3.28)	(3.87)	(3.79)	(2.32)	(2.16)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Org-Comzone FE	Yes	No	Yes	No	Yes	No
Industry×Stage×Year FE	Yes	No	Yes	No	Yes	No
Org-Comzone×Industry×Stage×Year FE	No	Yes	No	Yes	No	Yes
R-squared	0.16	0.21	0.31	0.33	0.26	0.37
N	76795	72542	78440	74187	39259	35205

Table 4: **Data Technology and Areas with Low History of VC Activity - IV Approach**

This table reports results for regressions for the instrumental variable two-stage least squares analysis at the VC-investment level, investigating whether the investments made by data-driven VCs after they adopt data technologies lead to different outcomes than those made by other VCs. Column (1) shows the first stage of the regression, where an indicator equal to one if an investment is made by a data-driven VC is fitted with the VC Exposure × Post measure. In columns (2) and (3), the dependent variable is an indicator if a VC made an investment in a startup located in a commuting zone in the lowest decile of VC activity. In columns (4) and (5) the dependent variable is an indicator if a VC made an investment in a startup located in a state in the lowest decile of VC activity. Columns (2) and (4) show the OLS results for the reduced sample. Columns (3) and (5) show the 2SLS results. All columns include VC-firm fixed effects, startup commuting zone fixed effects and startup industry by funding stage by funding year fixed effects. Control variables include the logarithm of the age of the VC firm and the logarithm of the startup's age. Regressions are double clustered at the VC firm year level.

Outcomes:		$\mathbb{1}$(Low VC Activity CZ)		$\mathbb{1}$(Low VC Activity State)	
	First Stage	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)
Data-Driven		0.015^{***}	0.196^{***}	0.025^{***}	0.174^{***}
		(4.19)	(2.61)	(3.08)	(2.43)
VC Exposure × Post	0.365^{***}				
	(2.79)				
Log(Startup Age)	-0.003	-0.003	-0.002	-0.001	-0.001
	(-1.45)	(-1.19)	(-0.97)	(-0.43)	(-0.27)
Log(VC Firm Age)	-0.021	-0.014^{**}	-0.010	-0.014^{**}	-0.010
	(-0.61)	(-2.37)	(-1.14)	(-2.18)	(-1.19)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
Org-Comzone FE	Yes	Yes	Yes	Yes	Yes
Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes
F-Statistic	12.45				
R-squared		0.79	-0.04	0.67	-0.02
N	44683	44683	44683	44683	44683

Table 5: **Data Technology and Other Measures of Out-of-Network Investments - IV Approach**

This table reports results for regressions for the instrumental variable two-stage least squares analysis at the VC-investment level, investigating whether the investments made by data-driven VCs after they adopt data technologies lead to different outcomes than those made by other VCs. Column (1) shows the first stage of the regression, where an indicator equal to one if an investment is made by a data-driven VC is fitted with the VC Exposure × Post measure. In columns (2) and (3), the dependent variable is an indicator if a VC made an investment in a startup located in the top tercile of distance from the VCs' headquarters. In columns (4) and (5), the dependent variable is an indicator if a VC made an investment in a a different industry than their specialization. In columns (6) and (7), the dependent variable is an indicator if a VC made an investment without a local syndicate conditinal on investing out of the VCs' headquarter state. Columns (2), (4) and (6) show the OLS results for the reduced sample. Columns (3), (5) and (7) show the 2SLS results. All columns include VC-firm fixed effects startup commuting zone by startup industry by funding stage by funding year fixed effects. Control variables include the logarithm of the age of the VC firm and the logarithm of the startup's age. Regressions are double clustered at the VC firm year level.

Outcomes:		$\mathbb{1}$(Top Tercile Distance)		$\mathbb{1}$(Diff Industry)		$\mathbb{1}$(Local Syndicate)	
	First Stage	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Data-Driven		0.025**	0.361**	0.108**	0.120***	-0.042*	-0.295
		(1.95)	(2.10)	(2.55)	(2.34)	(-1.40)	(-1.02)
VC Exposure × Post	0.408***						
	(2.98)						
Log(Startup Age)	-0.004	0.008*	0.010**	0.001	0.001	-0.063***	-0.062***
	(-1.53)	(2.05)	(2.15)	(0.34)	(0.30)	(-8.05)	(-7.81)
Log(VC Firm Age)	-0.019	0.023*	0.031*	-0.030	-0.029	0.018	0.026
	(-0.53)	(1.85)	(1.79)	(-0.95)	(-0.89)	(1.03)	(1.28)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Org-Comzone×Industry×Stage×Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Statistic	13.68						
R-squared		0.23	-0.01	0.40	0.00	0.36	-0.01
N	39946	39946	39946	40814	40814	18099	18099

Table 6: **Data-Driven Investment Entry and Entrepreneurial Activity**

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

Outcomes:	#First VC Financing		# VC Patents		# Entrep Patents	
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: All Controls			
Treat×Post	0.285***	0.220**	0.314**	0.309**	0.046	0.037
	(2.99)	(2.27)	(2.45)	(2.40)	(0.23)	(0.22)
VC Funding×Post	-0.282***	-0.430***	0.015	0.008	0.092	-0.039
	(-6.76)	(-7.65)	(0.12)	(0.06)	(0.80)	(-0.19)
Income×Post		0.084		0.479		0.727
		(0.32)		(0.89)		(1.40)
GDP×Post		0.178***		-0.040		0.144
		(3.68)		(-0.39)		(0.91)
Perc College×Post		2.231***		0.570		0.966
		(2.76)		(0.49)		(0.52)
Cohort×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×Commuting Zone FE	Yes	Yes	Yes	Yes	Yes	Yes
N	47524	46764	44581	43866	44581	43866
			Panel B: Controls with VC Activity			
Treat×Post	0.245**	0.175*	0.284	0.297	0.052	0.078
	(2.37)	(1.74)	(1.27)	(1.24)	(0.26)	(0.49)
VC Funding×Post	0.014	-0.186***	0.039	0.038	0.316*	0.149
	(0.24)	(-2.76)	(0.28)	(0.24)	(1.71)	(0.66)
Income×Post		-0.145		0.998		1.466*
		(-0.45)		(1.45)		(1.81)
GDP×Post		0.228***		-0.087		0.150
		(3.75)		(-0.78)		(0.70)
Perc College×Post		2.857***		0.097		0.259
		(2.98)		(0.07)		(0.09)
Cohort×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×Commuting Zone FE	Yes	Yes	Yes	Yes	Yes	Yes
N	12771	12421	11930	11604	11930	11604

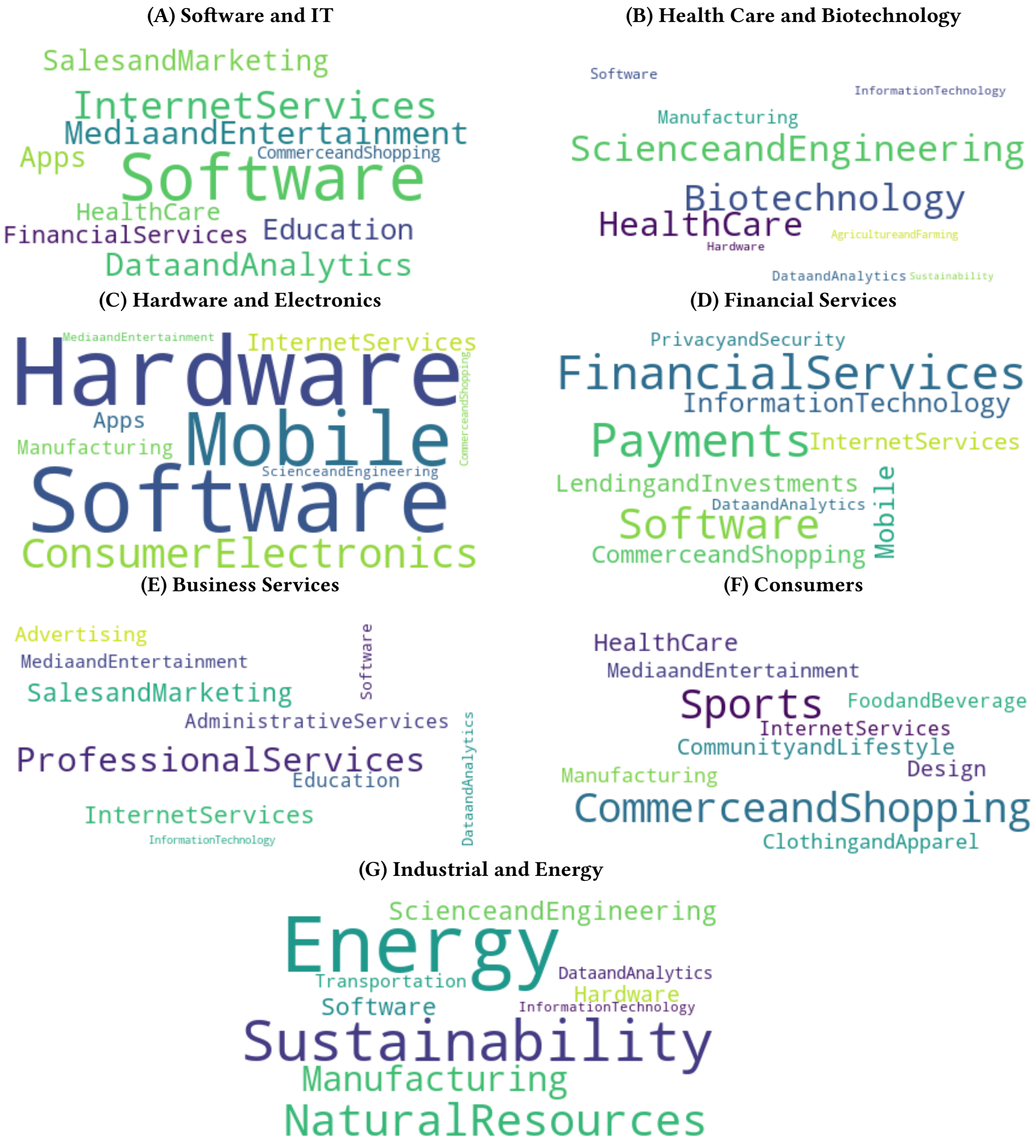
Table 7: **Data-Driven Investment Entry and VC Activity**

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

Outcomes:	#Funding Rounds		# Investors		# Investors First	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: All Controls						
Treat×Post	0.140**	0.118**	0.291***	0.234**	0.542***	0.449***
	(2.36)	(2.08)	(3.13)	(2.39)	(3.08)	(2.80)
VC Funding×Post	-0.161***	-0.239***	-0.361***	-0.544***	-0.247***	-0.449***
	(-5.91)	(-6.32)	(-7.55)	(-8.67)	(-4.64)	(-6.31)
Income×Post		0.100		0.178		0.357
		(0.59)		(0.55)		(0.99)
GDP×Post		0.093**		0.179***		0.184**
		(2.50)		(3.10)		(2.45)
Perc College×Post		1.020**		2.719***		3.086***
		(1.97)		(3.05)		(3.12)
Cohort×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×Commuting Zone FE	Yes	Yes	Yes	Yes	Yes	Yes
N	47524	46764	47524	46764	47524	46764
Panel B: Controls with VC Activity						
Treat×Post	0.112*	0.089	0.252**	0.196*	0.585**	0.472**
	(1.79)	(1.51)	(2.38)	(1.87)	(2.50)	(2.36)
VC Funding×Post	-0.102**	-0.190***	-0.116*	-0.348***	0.053	-0.203**
	(-2.55)	(-4.19)	(-1.78)	(-4.42)	(0.67)	(-2.17)
Income×Post		0.161		0.016		0.241
		(0.91)		(0.05)		(0.61)
GDP×Post		0.102**		0.230***		0.229**
		(2.34)		(3.32)		(2.49)
Perc College×Post		1.103*		2.987***		3.480***
		(1.89)		(2.95)		(3.06)
Cohort×Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort×Commuting Zone FE	Yes	Yes	Yes	Yes	Yes	Yes
N	12771	12421	12771	12421	12771	12421

Appendix

Figure A1: **Industry Classifications**



(A) Software and IT

(B) Health Care and Biotechnology

(C) Hardware and Electronics

(D) Financial Services

(E) Business Services

(F) Consumers

(G) Industrial and Energy

Table A1: **Data Technology and Areas with Low History of VC Activity - Pre-Covid**

This table reports results for regressions at the VC-investment level prior to 2020, investigating whether the investments made by data-driven VCs after they adopt data technologies lead to different outcomes than those made by other VCs. In columns (1) through (3), the dependent variable is an indicator if a VC made an investment in a startup located in a commuting zone in the lowest decile of VC activity. In columns (4) through (6), the dependent variable is an indicator if a VC made an investment in a startup located in a state in the lowest decile of VC activity. All columns include startup commuting zone fixed effects and startup industry by funding stage by funding year fixed effects. Columns (2), (3), (5), and (6) include VC firm fixed effects. Control variables across all specifications include the logarithm of the age of the VC firm and the logarithm of the startup's age. Control variables in columns (3) and (6) include the natural log of startup county's GDP and income and the percentage of the population that received a college degree a year prior to the investment. Regressions are double clustered at the VC firm year level.

Outcomes:	$\mathbb{1}$(Low VC Activity CZ)			$\mathbb{1}$(Low VC Activity State)		
	(1)	(2)	(3)	(4)	(5)	(6)
Data-Driven	0.009**	0.011***	0.007***	0.015**	0.026**	0.012*
	(2.31)	(3.86)	(3.36)	(2.57)	(2.75)	(1.78)
Log(Startup Age)	-0.001	-0.001	-0.001	-0.003	-0.001	-0.000
	(-1.07)	(-0.65)	(-0.49)	(-1.10)	(-0.63)	(-0.04)
Log(VC Firm Age)	-0.002*	-0.009**	-0.005*	-0.002**	-0.004	-0.001
	(-2.02)	(-2.23)	(-1.85)	(-2.14)	(-0.79)	(-0.16)
Log(GDP)			0.297***			0.292*
			(5.44)			(1.90)
Log(Income)			0.016			0.687*
			(0.12)			(1.93)
Perc College			-1.063***			0.775
			(-2.90)			(1.30)
VC-Firm FE	No	Yes	Yes	No	Yes	Yes
Org-Comzone FE	Yes	Yes	Yes	Yes	Yes	Yes
IndustryXStageXYear FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.81	0.82	0.86	0.71	0.72	0.75
N	57145	57093	55435	57145	57093	55435

Table A2: **Data Technology and Other Measures of Out-of-Network Investments**

This table reports results for regressions at the VC-investment level before 2020, investigating whether the investments made by data-driven VCs after they adopt data technologies lead to different outcomes than those made by other VCs. In columns (1) through (2), the dependent variable is an indicator if a VC made an investment in a startup located in the top tercile of distance from their headquarters. In columns (3) and (4), the dependent variable is an indicator if a VC invested in a startup in a different industry from their specialization. In columns (5) and (6), the dependent variable is an indicator if a VC invested in a startup without a local VC syndicate, conditional on investing out of their headquartered state. All columns include VC firm fixed effects. Odd columns include startup commuting zone fixed effects and startup industry by funding stage by funding year fixed effects. Even columns include startup commuting zone by startup industry by funding stage by funding year fixed effects. Control variables include the logarithm of the age of the VC firm and the logarithm of the startup's age. Regressions are double clustered at the VC firm year level.

Outcomes:	$\mathbb{1}$(Top Tercile Distance)		$\mathbb{1}$(Diff Industry)		$\mathbb{1}$(Local Syndicate)	
	(1)	(2)	(3)	(4)	(5)	(6)
Data-Driven	0.012**	0.005	0.070**	0.079*	-0.018	-0.049
	(2.59)	(1.21)	(2.44)	(1.81)	(-1.60)	(-1.50)
Log(Startup Age)	0.005	0.007*	-0.003	-0.002	-0.056***	-0.066***
	(1.19)	(1.74)	(-1.06)	(-0.63)	(-12.58)	(-12.18)
Log(VC Firm Age)	0.024**	0.021**	0.077***	0.074***	0.018	0.009
	(2.64)	(2.56)	(4.06)	(3.99)	(1.46)	(0.68)
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Org-Comzone FE	Yes	No	Yes	No	Yes	No
Industry×Stage×Year FE	Yes	No	Yes	No	Yes	No
Org-Comzone×Industry×Stage×Year FE	No	Yes	No	Yes	No	Yes
R-squared	0.20	0.25	0.33	0.35	0.27	0.38
N	55753	52361	57093	53696	26547	23336

Table A3: **AI Glossary**

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

Table A4: **Industry Exposure to AI**

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