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

Melissa Crumling
Drexel University

November 11, 2024

Motivation

- **Information acquisition is necessary for decision-making in financial markets**

Motivation

- **Information acquisition is necessary for decision-making in financial markets**
  - However, acquiring information can be costly

Motivation

- **Information acquisition is necessary for decision-making in financial markets**
  - However, acquiring information can be costly
- Various mechanisms to reduce information barriers
  - Geographic proximity (e.g. Van Nieuwerburgh and Veldkamp (2009))
  - Social networks (e.g. Kuchler et al. (2022))

Motivation

- **Information acquisition is necessary for decision-making in financial markets**
  - However, acquiring information can be costly
- Various mechanisms to reduce information barriers
  - Geographic proximity (e.g. Van Nieuwerburgh and Veldkamp (2009))
  - Social Networks (e.g. Kuchler et al. (2022))
  - **This paper:** the role of advanced data technologies

Motivation

- **Information acquisition is necessary for decision-making in financial markets**
  - However, acquiring information can be costly
- Various mechanisms to reduce information barriers
  - Geographic proximity (e.g. Van Nieuwerburgh and Veldkamp (2009))
  - Social Networks (e.g. Kuchler et al. (2022))
  - **This paper:** the role of advanced data technologies
- **RQ:** Do data technologies $\downarrow$ information frictions for financial decision making?

Motivation

- **Information acquisition is necessary for decision-making in financial markets**
  - However, acquiring information can be costly
- Various mechanisms to reduce information barriers
  - Geographic proximity (e.g. Van Nieuwerburgh and Veldkamp (2009))
  - Social Networks (e.g. Kuchler et al. (2022))
  - **This paper:** the role of advanced data technologies
- **RQ:** Do data technologies $\downarrow$ information frictions for financial decision making?
  - Does this impact who receives financing?

Motivation

- **Information acquisition is necessary for decision-making in financial markets**
    - However, acquiring information can be costly
- Various mechanisms to reduce information barriers
    - Geographic proximity (e.g. Van Nieuwerburgh and Veldkamp (2009))
    - Social Networks (e.g. Kuchler et al. (2022))
    - **This paper:** the role of advanced data technologies
- **RQ:** Do data technologies $\downarrow$ information frictions for financial decision making?
    - Does this impact who receives financing?
- Use the **Venture Capital** (VC) industry as a laboratory

Venture Capitalists as a Laboratory

- **VCs:** gatekeepers for high-growth startup financing
- Increasingly adopting data technologies to aid in investment process
- **VC industry provides an insightful setting:**
  - Important capital providers $\sim$ 50% of public firms VC-backed (e.g. Gornall and Strebulaev (2021))
    - Information frictions salient, significant shift from traditional approaches
  - Use of Big Data to inform investment decisions
    - 5 Vs: volume, velocity, variety, veracity, value

Case Study: Lightspeed Venture Partners

Founded: 2000

Case Study: Lightspeed Venture Partners

Founded: 2000

Hired First Data Scientist: Sep 2018

RAVI MHATRE BARRY EGGERS PETER NIEH

San Jose San Francisco New York

JERRY YE

"I help with widening the aperture so that we can algorithmically see more of what is happening in the startup ecosystem. **Data on company performance is generated every second. I extort signal from data.**"

This Paper: Do Data Technologies $\downarrow$ information frictions?

Methodology:

- Identify VC firms as data-driven from date of hiring first data-driven employee

This Paper: Do Data Technologies $\downarrow$ information frictions?

Methodology:

- Identify VC firms as data-driven from date of hiring first data-driven employee

Empirical Strategy:

- Use geographic makeup of VC industry
- VC activity concentrated in three main areas $\Rightarrow$ CA, MA, and NY
    - VC funds received 85% of capital raised (NVCA, 2020)
    - Startups received 73% of capital invested (NVCA, 2020)
- Startups not located in these areas likely fall outside of traditional VC networks
    - Examine where VCs choose to invest before and after technology adoption

▸ investment process

Main Hypothesis

H1$_a$: Data technologies $\downarrow$ information frictions for finding investment opportunities

H1$_0$: Data technologies have *limited* impact on information frictions

Main Hypothesis

H1$_a$: Data technologies $\downarrow$ information frictions for finding investment opportunities

- Broader discovery of startups beyond traditional networks
- Tracking real-time market trends and competitive dynamics
- Systematic approach for filtering out less promising startups

H1$_0$: Data technologies have *limited* impact on information frictions

Main Hypothesis

H1$_a$: Data technologies $\downarrow$ information frictions for finding investment opportunities

- Broader discovery of startups beyond traditional networks
- Tracking real-time market trends and competitive dynamics
- Systematic approach for filtering out less promising startups

H1$_0$: Data technologies have *limited* impact on information frictions

- Geographic separation remains a significant barrier to effective monitoring
- Data gaps and limitations
- Best startups located in hub areas

Case Study Revisited: Lightspeed Venture Partners

$$X = \text{DataDriven}_{j,t} = \begin{cases} 1, & \text{if Year}_t > 2018 \\ 0, & \text{otherwise} \end{cases}$$

Data Scientist Hire

	Jerry Ye	2018	Data Platform
	Tin Kyaw	2019	VP, Data
	Eric Wayman	2019	Staff Data Scientist
	Len Frenkel	2021	Software Engineer
	Radhika M.	2023	Data Scientist

Case Study Revisited: Lightspeed Venture Partners

$$Y = \#\text{Investments}_{j,t} \mid \text{Hub or Non-Hub} \qquad X = \text{DataDriven}_{j,t} = \begin{cases} 1, & \text{if Year}_t > 2018 \\ 0, & \text{otherwise} \end{cases}$$

First Time Investments			
2015-2018		**2019-2022**	
San Francisco, CA	66	San Francisco, CA	75
San Jose, CA	27	San Jose, CA	22
NY	17	NY	29
MA	3	MA	7
IL	1	IL	2
NJ	1	NJ	2
TX	1	TX	4
UT	1	UT	1
WA	1	WA	7
		DC	1
		DE	1
		FL	4
		GA	1
		MI	1
		NC	1
		NM	1
		NV	1

Data Scientist Hire			
	Jerry Ye	2018	Data Platform
	Tin Kyaw	2019	VP, Data
	Eric Wayman	2019	Staff Data Scientist
	Len Frenkel	2021	Software Engineer
	Radhika M.	2023	Data Scientist

Case Study Revisited: Lightspeed Venture Partners

$$Y = \#\text{Investments}_{j,t} \mid \text{Hub or Non-Hub} \qquad X = \text{DataDriven}_{j,t} = \begin{cases} 1, & \text{if Year}_t > 2018 \\ 0, & \text{otherwise} \end{cases}$$

First Time Investments					
2015-2018			2019-2022		
San Francisco, CA		66	San Francisco, CA		75
San Jose, CA		27	San Jose, CA		22
NY		17	NY		29
MA		3	MA		7
IL		1	IL		2
NJ		1	NJ		2
TX	4%	1	TX	17%	4
UT		1	UT		1
WA		1	WA		7
			DC		1
			DE		1
			FL		4
			GA		1
			MI		1
			NC		1
			NM		1
			NV		1

Data Scientist Hire

	Jerry Ye	2018	Data Platform
	Tin Kyaw	2019	VP, Data
	Eric Wayman	2019	Staff Data Scientist
	Len Frenkel	2021	Software Engineer
	Radhika M.	2023	Data Scientist

Contribution

Data Technologies and Pre-Investment Screening in VC Industry

Fintech and Information Production in Broader Financial Markets

Contribution

Data Technologies and Pre-Investment Screening in VC Industry

- Man vs Machine Retterath (2020), Lyonnet & Stern (2022), Davenport (2022) **This Paper: Ex post, VCs invest in more geographically diverse regions**
- Do not provide advantages for identifying "home run" investments Bonelli (2023) **This Paper: Heterogeneity depending on location of startup**

Fintech and Information Production in Broader Financial Markets

Contribution

Data Technologies and Pre-Investment Screening in VC Industry

- Man vs Machine Retterath (2020), Lyonnet & Stern (2022), Davenport (2022) **This Paper: Ex post, VCs invest in more geographically diverse regions**
- Do not provide advantages for identifying "home run" investments Bonelli (2023) **This Paper: Heterogeneity depending on location of startup**

Fintech and Information Production in Broader Financial Markets

- Market informativeness Weller (2018), Goa & Huang (2020), Abis (2022), Abis & Veldkamp (2022) **This Paper: Identify otherwise overlooked investments**
- Real Effects Zhu (2019), Bird, Karolyi, Ruchti & Truong (2021), Cao, Jiang, Yang & Zhang (2022), Dessaint, Foucault, & Fresard (2022), Goldstein, Yang & Zhou (2022) **This Paper: Increased VC activity outside traditional hubs**

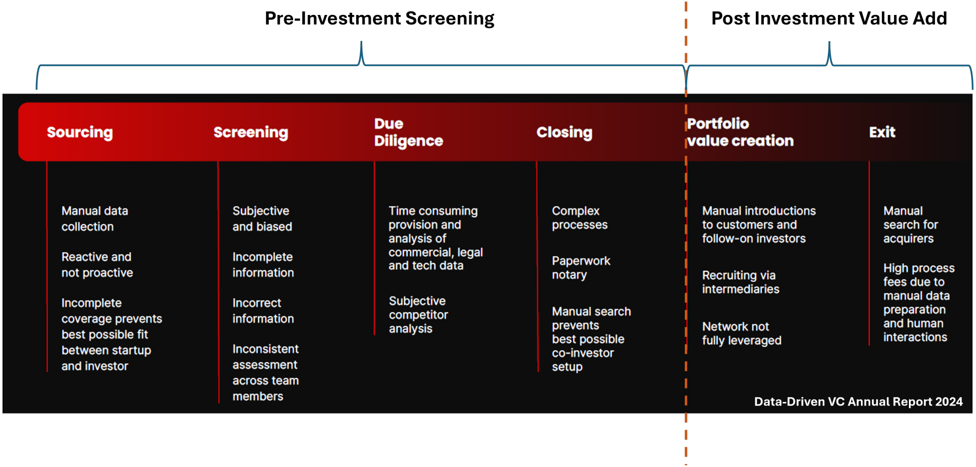
Roadmap and Overview of Findings

1. Background
2. Data and Methodology
3. Do Data Technologies $\downarrow$ information frictions for finding investment opportunities?
   - Do VCs $\uparrow$ # investments in non hub and low activity locations? **Find:** Yes
   - What are the main endogeneity concerns? **Firm Growth:** Conduct placebo analysis
4. How do Data-Driven Non Hub Investments Perform?
   Find: More likely to IPO than 1) DD Inv in Hubs & 2) Trad Inv in Non Hubs
5. Do these areas experience an $\uparrow$ in subsequent VC activity? **Find:** Yes

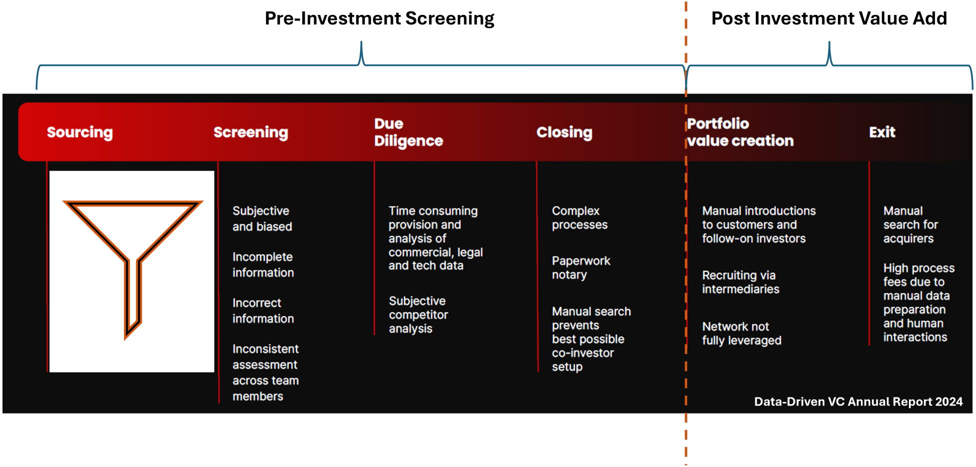
Roadmap and Overview of Findings

1. Background
2. Data and Methodology
3. Do Data Technologies $\downarrow$ information frictions for finding investment opportunities?
   - Do VCs $\uparrow$ # investments in non hub and low activity locations?
   - What are the main endogeneity concerns?
4. How do Data-Driven Non Hub Investments Perform?
5. Do these areas experience an $\uparrow$ in subsequent VC activity?

VC Investment Process



VC Investment Process



Sourcing Investments

Table 3
Sources of investments.

	All
Inbound from management	10
	(1)
Referred by portfolio company	8
	(1)
Referred by other investors	20
	(1)
Professional network	31
	(1)
Proactively self-generated	28
	(1)
Quantitative sourcing	2
	(0)
Number of responses	446

Source: Gompers et al. (2020)

~60% Network

Lerner (1994), Hochberg et al. (2007), Hochberg et al. (2010), Gompers et al. (2016), Garfinkel et al. (2024), Huang (2024)

Sourcing Investments

Table 3
Sources of investments.

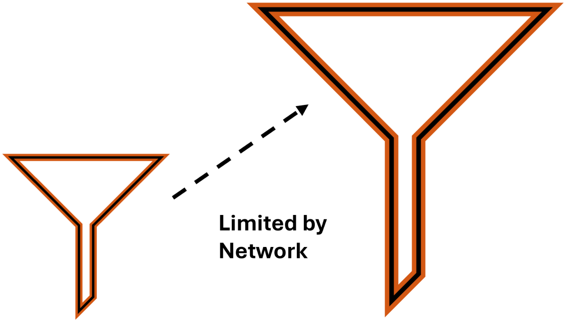
	All
Inbound from management	10
	(1)
Referred by portfolio company	8
	(1)
Referred by other investors	20
	(1)
Professional network	31
	(1)
Proactively self-generated	28
	(1)
Quantitative sourcing	2
	(0)
Number of responses	446

Source: Gompers et al. (2020)

~60% Network

Lerner (1994), Hochberg et al. (2007), Hochberg et al. (2010), Gompers et al. (2016), Garfinkel et al. (2024), Huang (2024)

"Above all, VC is a network business, effectively capped by the scalability of human relationships" – Damian Cristian, Koble VC

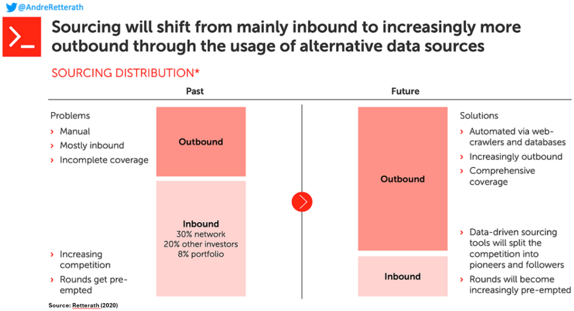


Data-Driven Approaches

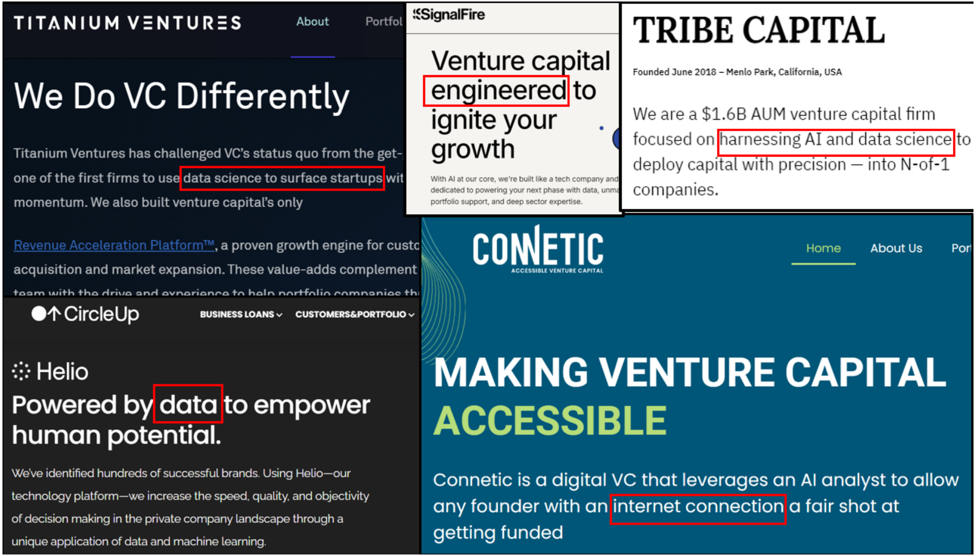
- Use webcrawlers and alternative data to identify startups independent of location
  - e.g. Github, public registers, LinkedIn, Twitter
- Once identified, enrich to create a comprehensive picture of company
  - e.g. Crunchbase, Pitchbook, website traffic, App Store info

Data-Driven Approaches

- Use webcrawlers and alternative data to identify startups independent of location
  - e.g. Github, public registers, LinkedIn, Twitter
- Once identified, enrich to create a comprehensive picture of company
  - e.g. Crunchbase, Pitchbook, website traffic, App Store info



Data-Driven VC Examples



Roadmap and Overview of Findings

1. Background
2. Data and Methodology
3. Do Data Technologies $\downarrow$ information frictions for finding investment opportunities?
   - Do VCs $\uparrow$ # investments in non hub and low activity locations?
   - What are the main endogeneity concerns?
4. How do Data-Driven Non Hub Investments Perform?
5. Do these areas experience an $\uparrow$ in subsequent VC activity?

Data

- **VC Investments**
  - Crunchbase: keep all VCs headquartered in the US
    - Merge with Preqin and VentureXpert
  - 927 distinct VC firms from 2010 to 2022
  - Investment information, founding year, HQ location, industry, stage

- **Employee Histories**
  - Crunchbase and LinkedIn to find data-driven employees

- **Regional Entrepreneurial Activity**
  - Startup Cartography Project (Andrews, Fazio, Guzman, Liu and Stern (2019))
  - Entrepreneurial ecosystem statistics for US from 1988-2016
    - startup quantity and quality measures at the state, MSA, county and zip-code level

Identifying Data-Driven VCs

Identify VCs using data technologies as those who hire data-related employees

- Prior research used job postings to infer technology adoption (e.g. Bonelli (2023), Raymond (2024))

1. Identify initial list from Data-Driven VC (Retterath, Early Bird Ventures (2024)) $\rightarrow$ create job title list
2. Use job title list to identify data-driven VCs in my sample

Identifying Data-Driven VCs

Identify VCs using data technologies as those who hire data-related employees

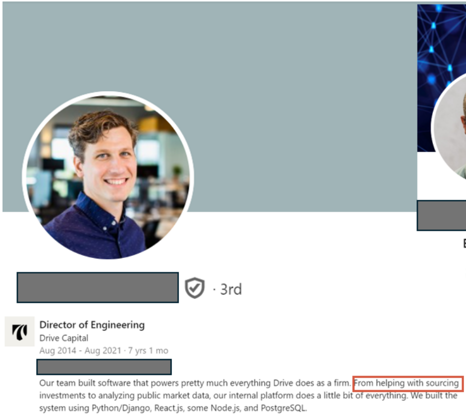
- Prior research used job postings to infer technology adoption (e.g. Bonelli (2023), Raymond (2024))

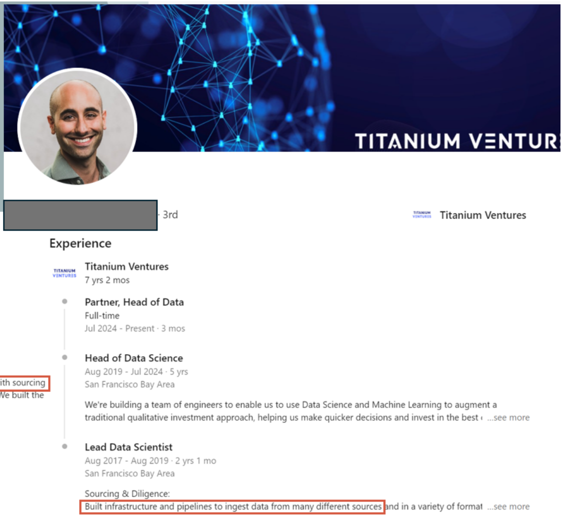
1. Identify initial list from Data-Driven VC (Retterath, Early Bird Ventures (2024)) $\rightarrow$ create job title list
2. Use job title list to identify data-driven VCs in my sample

$\Rightarrow$ 59 data-driven VCs from 2010 to 2022 $\rightarrow$ 2,965 data-driven investments

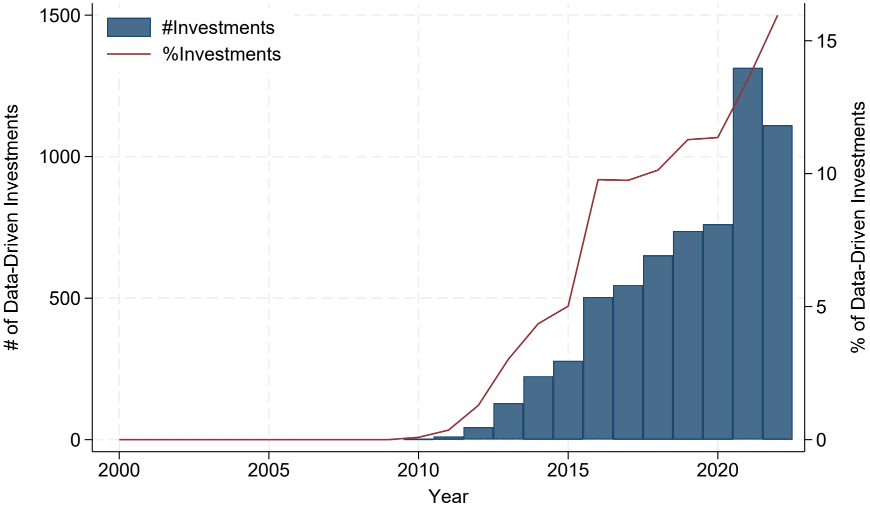
	Data-Driven		Traditional		Difference
	Mean	Count	Mean	Count	
Age	14.65	598	11.92	7915	2.73***
# Employees	23.18	598	8.95	7915	15.17***
AUM ($ Bil)	1.26	598	0.44	7915	0.82***
Centrality	5.93	598	2.69	7915	3.23***
Hub HQ	0.94	598	0.79	7915	0.15***
Software Industry	0.91	598	0.62	7915	0.29***

Data Scientist Examples





Evolution of Data-Driven VCs



Geographic Classifications

- **Hub**
  - commuting zones in San Francisco and San Jose, CA; Boston, MA; New York, NY
- **Non Hub**
  - all other commuting zones
- **Low Activity**
  - commuting zones with < 25 VC investments in previous 5 years (Hochberg et al. (2010))

Roadmap and Overview of Findings

1. Background
2. Data and Methodology
3. Do Data Technologies $\downarrow$ information frictions for finding investment opportunities?
   - Do VCs $\uparrow$ # investments in non hub and low activity locations?
   - What are the main endogeneity concerns?
4. Do Data-Driven Non Hub Investments Outperform Hub Investments?
5. Do these areas experience an $\uparrow$ in subsequent VC activity?

Specification: Stacked Diff-in-Diff (Baker et al. (2022))

$$Y_{j,d,t} = \beta_1 \mathit{Treated}_{j,d} \times \mathit{Post}_{d,t} + X_{j,d,t} + \alpha_{j\times d} + \alpha_{d\times c\times i\times s\times t} + \epsilon_{j,d,t}$$

- $y_{j,t}$ = # Investments made by VC j in year t
- $\mathit{Treated}_{j,d}$ = indicator if VC j becomes data-driven, 0 otherwise
- $\mathit{Post}_{d,t}$ = indicator after data-driven event
- $X_{j,d,t}$ = time varying controls for VC j
  - VC firm age, # of employees, total AUM, eigenvector network centrality
- α_j = VC firm fixed effects
- $\gamma_{i\times c\times t\times s}$ = VC main industry i $\times$ state c of VC HQ $\times$ year t $\times$ VC main funding stage s FEs

Specification: Stacked Diff-in-Diff (Baker et al. (2022))

$$Y_{j,d,t} = \beta_1 \, Treated_{j,d} \times Post_{d,t} + X_{j,d,t} + \alpha_{j\times d} + \alpha_{d\times c\times i\times s\times t} + \epsilon_{j,d,t}$$

- $y_{j,t}$ = # Investments made by VC j in year t
- $Treated_{j,d}$ = indicator if VC j becomes data-driven, 0 otherwise
- $Post_{d,t}$ = indicator after data-driven event
- $X_{j,d,t}$ = time varying controls for VC j
  - VC firm age, # of employees, total AUM, eigenvector network centrality
- α_j = VC firm fixed effects
- $\gamma_{i\times c\times t\times s}$ = VC main industry i $\times$ state c of VC HQ $\times$ year t $\times$ VC main funding stage s FEs

Prediction: $\beta_1 > 0$ – After adopting data technologies, VCs $\uparrow$ # investments in non-hubs

Main Results

$$\underbrace{y_{j,d,t}}_{\text{\# Investments}} = \beta \underbrace{Treat_j \times Post_t}_{\text{data-driven}} + \underbrace{X_{j,d,t}}_{\text{VC controls}} + \underbrace{\alpha_{j\times d}}_{\substack{\text{VC firm-}\\\text{by-cohort FE}}} + \underbrace{\gamma_{i\times c\times t\times s\times d}}_{\substack{\text{state-by-ind.-by-}\\\text{stage-by-year-by-cohort FE}}} + \epsilon_{j,t}$$

Outcomes:	Hub	Non Hub	Low Activity
	(1)	(2)	(3)
Data Driven	0.104	0.152***	0.460*
	(1.22)	(2.54)	(1.81)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State$\times$Industry$\times$Stage$\times$Year FE	Yes	Yes	Yes
$\bar{y}$	5.45	2.81	0.24
R-squared	0.51	0.71	0.51
N	31069	31069	31069

Main Results

$$\underbrace{y_{j,d,t}}_{\text{\# Investments}} = \beta \underbrace{Treat_j \times Post_t}_{\text{data-driven}} + \underbrace{X_{j,d,t}}_{\text{VC controls}} + \underbrace{\alpha_{j\times d}}_{\text{VC firm-by-cohort FE}} + \underbrace{\gamma_{i\times c\times t\times s\times d}}_{\text{state-by-ind.-by-stage-by-year-by-cohort FE}} + \epsilon_{j,t}$$

Outcomes:	Hub	Non Hub	Low Activity
	(1)	(2)	(3)
Data Driven	0.104	0.152***	0.460*
	(1.22)	(2.54)	(1.81)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes
$\bar{y}$	5.45	2.81	0.24
R-squared	0.51	0.71	0.51
N	31069	31069	31069

- Investments in Non Hub Areas $\uparrow$ by $e^{0.152} - 1 = 16\% \implies \sim 0.5$ inv. per year
- Investments in Low Act. Areas $\uparrow$ by $e^{0.46} - 1 = 58\% \implies \sim 0.15$ inv. per year

Endogeneity Concerns

Hiring of Data Scientist correlated with overall firm growth ▸ fund & employee analysis

- Include # Employees and AUM as controls

Empirical Approach: conduct placebo analysis with hiring of a **Venture Partner** (VP)

- Distinct from General Partners (GP)
  - No carried interest, focus on sourcing, no investment authority
- Increasingly common as influx of capital to private markets (Razaei (2024))

Intuition: VCs hire when growing $\rightarrow$ correlated with increased investments

- Compare VP hire to data scientist hire to isolate unique impact of data technology

Approach

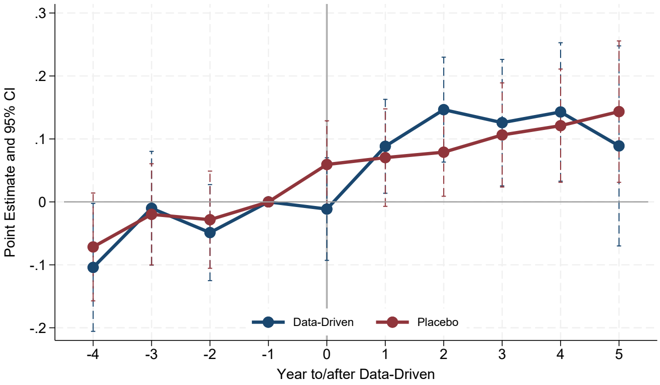
Match each VC that hires a data scientist to a VC that hires a VP in the same year

- On covariates and pre-trends
  - Age, # Employees, AUM, Centrality
  - State, Industry, Stage

	DD Hire		VP Hire		Difference
	Mean	**N**	**Mean**	**N**	**DD-VP**
Age	12.44	398	12.65	358	-0.21
# Employees	22.24	398	19.01	358	3.23**
AUM	1.35	398	1.16	358	0.19
Centrality	5.08	398	5.13	358	-0.05

Total Investment with Placebo

$$\underbrace{y_{j,d,t}}_{\text{\# Investments}} = \sum_{k=-4,k\neq-1}^{5} \beta \underbrace{Treat_{j,d}}_{\substack{\text{data-driven}\\ \text{or VP}}} \times Year(k)_d + \underbrace{X_{j,d,t}}_{\text{VC controls}} + \underbrace{\alpha_{j\times d}}_{\substack{\text{VC firm-}\\ \text{by-cohort FE}}} + \underbrace{\gamma_{i\times c\times t\times s\times d}}_{\substack{\text{state-by-ind.-by-}\\ \text{stage-by-year-by-cohort FE}}} + \epsilon_{j,t}$$



Main Results with Placebo

$$\underbrace{y_{j,d,t}}_{\text{\# Inv.}} = \beta_1 \underbrace{DataDriven_j \times Post_t}_{\text{data-driven}} + \beta_2 \underbrace{Placebo_j \times Post_t}_{\text{VP Hire}} + \underbrace{X_{j,d,t}}_{\text{VC controls}} + \underbrace{\alpha_{j\times d}}_{\text{VC firm-by-cohort FE}} + \underbrace{\gamma_{i\times c\times t\times s\times d}}_{\text{state-by-ind.-by-stage-by-year-by-cohort FE}} + \epsilon_{j,t}$$

Outcomes:	Hub	Non Hub	Low Activity
	(1)	(2)	(3)
Data Driven$\times$Post	0.112	0.167***	0.466**
	(0.99)	(2.66)	(2.17)
Placebo$\times$Post	0.134**	0.096	-0.037
	(1.98)	(0.94)	(-0.18)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State$\times$Industry$\times$Stage$\times$Year FE	Yes	Yes	Yes
Data Driven$\times$Post = Placebo$\times$Post (p-value)	0.865	0.543	0.0406**
$\bar{y}$	5.45	2.81	0.24
R-squared	0.51	0.71	0.51
N	30855	30855	30855

▸ treat dynamics ▸ placebo dynamics

Main Results with Placebo

$$\underbrace{y_{j,d,t}}_{\text{\# Inv.}} = \beta_1 \underbrace{DataDriven_j \times Post_t}_{\text{data-driven}} + \beta_2 \underbrace{Placebo_j \times Post_t}_{\text{VP Hire}} + \underbrace{X_{j,d,t}}_{\text{VC controls}} + \underbrace{\alpha_{j\times d}}_{\text{VC firm-by-cohort FE}} + \underbrace{\gamma_{i\times c\times t\times s\times d}}_{\text{state-by-ind.-by-stage-by-year-by-cohort FE}} + \epsilon_{j,t}$$

Outcomes:	Hub	Non Hub	Low Activity
	(1)	(2)	(3)
Data Driven×Post	0.112	0.167***	0.466**
	(0.99)	(2.66)	(2.17)
Placebo×Post	0.134**	0.096	-0.037
	(1.98)	(0.94)	(-0.18)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes
Data Driven×Post = Placebo×Post (p-value)	0.865	0.543	0.0406**
$\bar{y}$	5.45	2.81	0.24
R-squared	0.51	0.71	0.51
N	30855	30855	30855

▸ treat dynamics ▸ placebo dynamics

Main Results with Placebo

$$\underbrace{y_{j,d,t}}_{\text{\# Inv.}} = \beta_1 \underbrace{DataDriven_j \times Post_t}_{\text{data-driven}} + \beta_2 \underbrace{Placebo_j \times Post_t}_{\text{VP Hire}} + \underbrace{X_{j,d,t}}_{\text{VC controls}} + \underbrace{\alpha_{j\times d}}_{\substack{\text{VC firm-}\\ \text{by-cohort FE}}} + \underbrace{\gamma_{i\times c\times t\times s\times d}}_{\substack{\text{state-by-ind.-by-}\\ \text{stage-by-year-by-cohort FE}}} + \epsilon_{j,t}$$

Outcomes:	Hub	Non Hub	Low Activity
	(1)	(2)	(3)
Data Driven $\times$ Post	0.112	0.167***	0.466**
	(0.99)	(2.66)	(2.17)
Placebo $\times$ Post	0.134**	0.096	-0.037
	(1.98)	(0.94)	(-0.18)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State $\times$ Industry $\times$ Stage $\times$ Year FE	Yes	Yes	Yes
Data Driven $\times$ Post $=$ Placebo $\times$ Post (p-value)	0.865	0.543	0.0406**
$\bar{y}$	5.45	2.81	0.24
R-squared	0.51	0.71	0.51
N	30855	30855	30855

▸ treat dynamics ▸ placebo dynamics

Additional Tests

- Other Measures of Data-Driven
  - $Log(1 + \#\text{ Data Scientists})$, $\frac{\#\text{Data Scientists}}{\#\text{ GPs}}$ ‣ other measures
- Do Data-Driven VCs invest in the same non hubs?
  - No, # of non hub commuting zones (states) $\uparrow$ by 21% (24%) ‣ num locations
- Other Proxies for Information Asymmetry ‣ other proxies
  - More likely to invest in different industry
  - Less likely to invest with local syndicate
  - More likely to lead funding round
- Selection - IV Approach ‣ IV approach
  - VCs pre-exposure to data technologies and timing of raising a new fund

Roadmap and Overview of Findings

1. Background
2. Data and Methodology
3. Do Data Technologies $\downarrow$ information frictions for finding investment opportunities?
   - Do VCs $\uparrow$ # investments in non hub and low activity locations? **Find:** Yes
   - What are the main endogeneity concerns? **Firm Growth:** Conduct placebo analysis
4. **How do Data-Driven Non Hub Investments Perform?**
5. Do these areas experience an $\uparrow$ in subsequent VC activity?

Data-Driven Performance

- **Recap**: After technology adoption, VCs $\uparrow$ investments in non-hub & low activity areas
- Receive majority of returns through (rare) liquidity event
  - IPO (5-10x) $\rightarrow$ 7%, Acquisition (1-5x) $\rightarrow$ 23%
  - Achieve Unicorn status $\rightarrow$ 10%
- Ex ante, performance of non-hub startups unclear
  - Less competition for high quality startups, higher hurdle rate (e.g. Chen et al. (2010))
  - Difficult to assess quality ex ante, unable to monitor as effectively (e.g. Cumming and Dai (2010))

▸ DD monitoring ▸ follow on

Non Hub Performance

$$\underbrace{y_{j,k,t}}_{\text{IPO, unicorn or acquisition}} = \beta \underbrace{\textit{Data Driven}_{j,t}}_{\text{data-driven investor}} + \underbrace{X_{j,k,t}}_{\text{VC \& startup controls}} + \underbrace{\alpha_j}_{\text{VC firm FE}} + \underbrace{\gamma_{i \times c \times t \times s}}_{\text{state-by-ind.-by-stage-by-year FE}} + \epsilon_{j,t}$$

Non Hub Performance

$$\underbrace{y_{j,k,t}}_{\substack{\text{IPO, unicorn}\\ \text{or acquisition}}} = \beta \underbrace{\textit{Data Driven}_{j,t}}_{\substack{\text{data-driven}\\ \text{investor}}} + \underbrace{X_{j,k,t}}_{\substack{\text{VC \&}\\ \text{startup controls}}} + \underbrace{\alpha_j}_{\text{VC firm FE}} + \underbrace{\gamma_{i\times c\times t\times s}}_{\text{state-by-ind.-by-stage-by-year FE}} + \epsilon_{j,t}$$

Outcomes:	Major Success		
	(1)	(2)	(3)
Data Driven × Non Hub		0.001	
		(0.973)	
Data Driven × Low Activity			0.061
			(0.74)
Data Driven	0.029	0.028	0.028
	(1.26)	(1.32)	(1.24)
Non Hub & Low Activity	No	Yes	Yes
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State × Industry × Stage × Year FE	Yes	Yes	Yes
Data Driven × Non Hub = Data Driven (p-value)		0.4130	
Data Driven × Low Activity = Data Driven (p-value)			0.6935
$\bar{y}$	0.33	0.33	0.33
$\bar{y}$ | Non Hub, Low Activity		0.28	0.23
R-squared	0.31	0.31	0.31
N	22428	22428	22428

Non Hub Performance

$$\underbrace{y_{j,k,t}}_{\text{IPO, unicorn or acquisition}} = \beta \underbrace{Data\ Driven_{j,t}}_{\text{data-driven investor}} + \underbrace{X_{j,k,t}}_{\text{VC \& startup controls}} + \underbrace{\alpha_j}_{\text{VC firm FE}} + \underbrace{\gamma_{i \times c \times t \times s}}_{\text{state-by-ind.-by-stage-by-year FE}} + \epsilon_{j,t}$$

Outcomes:	IPO or Unicorn		Acquisition	
	(1)	(2)	(3)	(4)
Data Driven $\times$ Non Hub	0.026**		-0.022	
	(2.25)		(-0.98)	
Data Driven $\times$ Low Activity		0.045*		0.013
		(1.87)		(0.14)
Data Driven	0.007	0.012	0.007	0.002
	(0.49)	(0.83)	(0.29)	(0.08)
Non Hub & Low Activity	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes
State $\times$ Industry $\times$ Stage $\times$ Year FE	Yes	Yes	Yes	Yes
Data Driven $\times$ Non Hub $=$ Data Driven (p-value)	0.0245**		0.4333	
Data Driven $\times$ Low Activity $=$ Data Driven (p-value)		0.0099***		0.9141
$\bar{y}$	0.12	0.12	0.23	0.23
$\bar{y}$ | Non Hub, Low Activity	0.09	0.04	0.21	0.20
R-squared	0.16	0.16	0.14	0.14
N	22428	22428	22428	22428

IPO or uni full   acquisition full

Non Hub Performance

$$\underbrace{y_{j,k,t}}_{\text{IPO, unicorn or acquisition}} = \beta \underbrace{\textit{Data Driven}_{j,t}}_{\text{data-driven investor}} + \underbrace{X_{j,k,t}}_{\text{VC \& startup controls}} + \underbrace{\alpha_j}_{\text{VC firm FE}} + \underbrace{\gamma_{i \times c \times t \times s}}_{\text{state-by-ind.-by-stage-by-year FE}} + \epsilon_{j,t}$$

Outcomes:	IPO or Unicorn		Acquisition	
	(1)	(2)	(3)	(4)
Data Driven×Non Hub	0.026**		-0.022	
	(2.25)		(-0.98)	
Data Driven×Low Activity		0.045*		0.013
		(1.87)		(0.14)
Non Hub or Low Activity	-0.019*	-0.054***	-0.018	-0.026
	(-1.79)	(-2.85)	(-1.30)	(-0.89)
Data Driven	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes
State×Industry×Stage×Year FE	Yes	Yes	Yes	Yes
Data Driven×Non Hub = Non Hub (p-value)	0.0077***		0.8917	
Data Driven×Low Activity = Low Activity (p-value)		0.1408		0.7090
$\bar{y}$	0.12	0.12	0.23	0.23
$\bar{y}$|Non Hub, Low Activity	0.09	0.04	0.21	0.20
R-squared	0.16	0.16	0.14	0.14
N	22428	22428	22428	22428

Roadmap and Overview of Findings

1. Background
2. Data and Methodology
3. Do Data Technologies $\downarrow$ information frictions for finding investment opportunities?
   - Do VCs $\uparrow$ # investments in non hub and low activity locations? **Find:** Yes
   - What are the main endogeneity concerns? **Firm Growth:** Conduct placebo analysis
4. How do Data-Driven Non Hub Investments Perform?
   Find: More likely to IPO than 1) DD Inv in Hubs & 2) Trad Inv in Non Hubs Yes
5. Do these areas experience an $\uparrow$ in subsequent VC activity?

Do data-driven investments in low activity areas lead to $\uparrow$ VC activity?

- Once VCs invest in low activity areas, these areas are more likely to become part of ...
  - Databases used by data-driven VCs
  - Traditional VC networks

- ***Prediction:*** Data-driven investments in low-activity hubs lead to $\uparrow$ VC activity

- Identify all comzones with < 25 VC investments in last 5 years from 2010 to 2022
  - Treated comzones - received funding by data-driven VC$\rightarrow$ 56 comzones
    - All other comzones - control

VC Activity

$$\underbrace{y_{d,c,t}}_{\substack{\text{vc activity}\\ \text{outcomes}}} = \beta \underbrace{\{Treated_{d,c} \times Post_{d,t}\}}_{\substack{\text{cz receive}\\ \text{data-driven investment}}} + \underbrace{X_{d,c,t-1}}_{\text{cz controls}} + \underbrace{\alpha_{d,c}}_{\substack{\text{cohort-by-cz}\\ \text{FE}}} + \underbrace{\alpha_{d,t}}_{\substack{\text{cohort-by-year}\\ \text{FE}}} + \epsilon_{d,c,t}$$

VC Activity

$$\underbrace{y_{d,c,t}}_{\text{vc activity outcomes}} = \beta \underbrace{\{Treated_{d,c} \times Post_{d,t}\}}_{\text{cz receive data-driven investment}} + \underbrace{X_{d,c,t-1}}_{\text{cz controls}} + \underbrace{\alpha_{d,c}}_{\text{cohort-by-cz FE}} + \underbrace{\alpha_{d,t}}_{\text{cohort-by-year FE}} + \epsilon_{d,c,t}$$

Outcomes:	# Funding Rounds	# First VC Financing	# Unique Investors	# First Investor	# VC Patents
	(1)	(2)	(3)	(4)	(5)
Treat×Post	0.077**	0.084**	0.186***	0.112**	0.282***
	(2.59)	(2.63)	(3.12)	(2.65)	(3.64)
Controls	Yes	Yes	Yes	Yes	Yes
Cohort×Year FE	Yes	Yes	Yes	Yes	Yes
Cohort×Commuting Zone FE	Yes	Yes	Yes	Yes	Yes
R-squared	0.75	0.64	0.76	0.72	0.68
$\bar{y}$	0.31	0.15	0.57	0.42	0.33
N	53548	53548	53548	53548	53548

▸ entrepreneurial czs

Conclusion

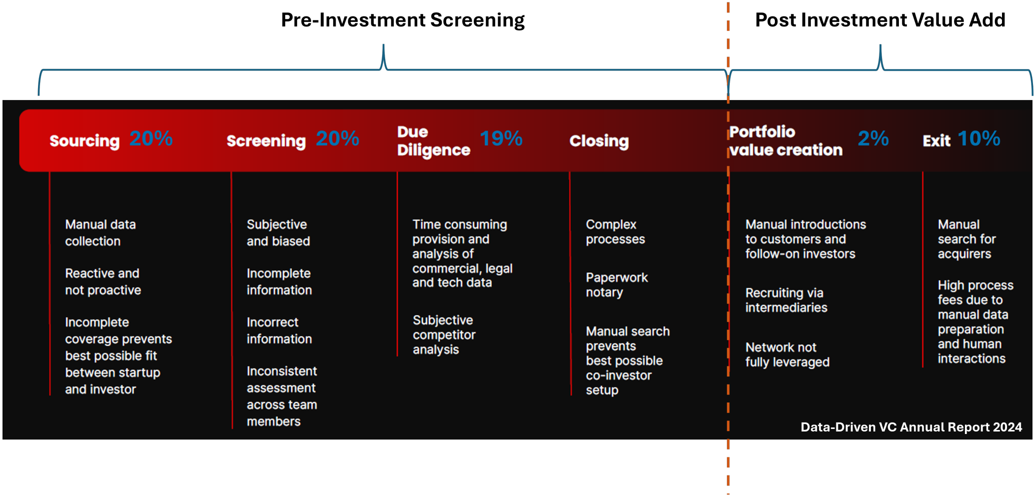
- **I study the use of data technologies to overcome information frictions**
  - Use the VC industry as a laboratory
- Data technologies $\downarrow$ search frictions, $\uparrow$ investments non hub commuting zones
  - These investments are more likely to exit through an IPO or achieve unicorn status
- Data-driven entry in low activity areas lead to an $\uparrow$ in subsequent VC activity
- Results suggest that data technologies change the way VCs source investments
  - Encourage regional innovation outside of traditional hubs

"Note to founders, start leaving your trails online about what you're building, if you're on the right path – they'll come knocking on your door" - Gabriel Shin, Landscape

Thank You!

Appendix

Investment Process & Data Driven Usage

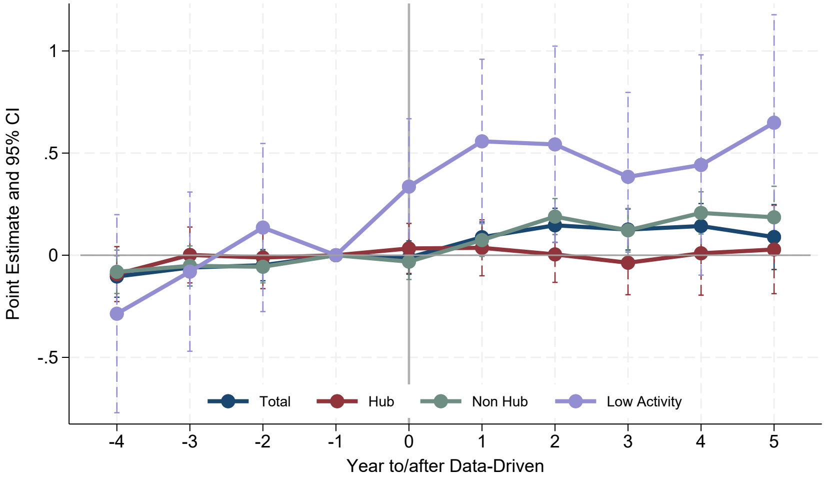


Firm Growth

Outcomes:	Fund Size		Employee Size	
	Log(Total AUM)	Log(Median Round $)	Log(# Partners)	# Inv/Partner
	(1)	(2)	(3)	(4)
Data Driven	0.247**	0.006	0.101*	0.030
	(2.39)	(0.08)	(2.14)	(0.28)
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes
State $\times$ Industry $\times$ Stage $\times$ Year FE	Yes	Yes	Yes	Yes
R-squared	0.90	0.69	0.87	0.85
N	8513	8513	8513	8513

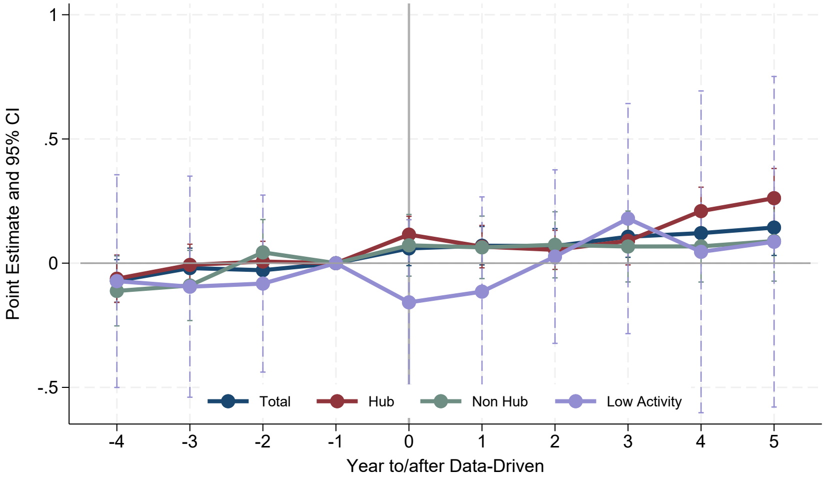
back

Dynamics - Treat



back

Dynamics - Placebo



back

Other Data Driven Measures

Data Driven =	$Log(1 + \#\text{ Data Scientists})$			$\frac{\#\text{Data Scientists}}{\#\text{ Partners}}$		
Outcomes:	Hub	Non Hub	Low	Hub	Non Hub	Low
	(1)	(2)	(3)	(4)	(5)	(6)
Data Driven	0.131	0.177***	0.389**	0.378	0.812***	1.218**
	(1.60)	(3.67)	(2.09)	(0.74)	(3.06)	(2.32)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
State$\times$Industry$\times$Stage$\times$Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.51	0.71	0.51	0.51	0.71	0.49
$\bar{y}$	5.45	2.81	0.24	5.45	2.81	0.24
N	8513	8513	8513	8513	8513	8513

back

Number Locations

Outcomes:	# Comzones	# Nonhub Comzones	# States	# Nonhub States
	(1)	(2)	(3)	(4)
Data Driven	0.156**	0.213**	0.129*	0.239**
	(2.25)	(1.98)	(1.91)	(2.52)
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes
State × Industry × Stage × Year FE	Yes	Yes	Yes	Yes
R-squared	0.26	0.29	0.22	0.28
N	8513	8513	8513	8513

back

Other Proxies for IA

Outcomes:	Diff Industry	Local Syndicate		Lead Investor	
	All	Non Hub	Low Activity	Non Hub	Low Activity
	(1)	(2)	(3)	(4)	(5)
Data-Driven	0.07*	-0.040**	-0.390***	0.064***	0.024
	(1.80)	(-2.16)	(-4.78)	(2.92)	(0.86)
Controls	Yes	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes	Yes
State × Industry × Stage × Year FE	Yes	Yes	Yes	Yes	Yes
R-squared	0.37	0.37	1.01	0.09	0.07
N	49411	6659	566	6659	566

back

IV Approach

VC's adoption of data technologies is not random ...

Correlated omitted variable with technology adoption and outcome

IV Strategy - isolate variation in VCs' data technology adoption from two sources:

1. early exposure to AI
2. timing of raising a new fund

▸ back

Identification Strategy [1]

Step 1: Exogenous variation in VCs' early exposure to AI

- Commercial interest in AI became widespread around 2010 Babina et al. (2024)
  - Tech firms - e.g. Apple introducing Siri in 2011
  - Non-tech firms - e.g. Walmart using cameras on floor scrubbers (2017)
- Startups some of the first to pioneer AI development in 2000s
  - e.g. Predictix, 2005; Voci, 2008
  - VCs that finance these startups have first mover advantage
  - Measure how much a VC is **exposed** to AI through its *investments* - before 2010

Identification Strategy [1]

Step 1: Exogenous variation in VCs' early exposure to AI

- Commercial interest in AI became widespread around 2010 Babina et al. (2024)
  - Tech firms - e.g. Apple introducing Siri in 2011
  - Non-tech firms - e.g. Walmart using cameras on floor scrubbers (2017)

- Startups some of the first to pioneer AI development in 2000s
  - e.g. Predictix, 2005; Voci, 2008
  - VCs that finance these startups have first mover advantage
  - Measure how much a VC is **exposed** to AI through its *investments* - before 2010

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

Identification Strategy [2]

Step 2: Timing of raising a new fund

- VCs typically hire new employees when raising a new fund
- Typically raise a fund every 3-5 years
  - Prior funds nearly deployed
  - External market conditions
- Therefore VCs are likely to hire a data scientist while fund raising
- $NewFund_{j,[-2:0]}$ indicates if VC raised a new fund in the previous 2 years

Identification Strategy [2]

Step 2: Timing of raising a new fund

- VCs typically hire new employees when raising a new fund
- Typically raise a fund every 3-5 years
  - Prior funds nearly deployed
  - External market conditions
- Therefore VCs are likely to hire a data scientist while fund raising
- $NewFund_{j,[-2:0]}$ indicates if VC raised a new fund in the previous 2 years

First Stage:

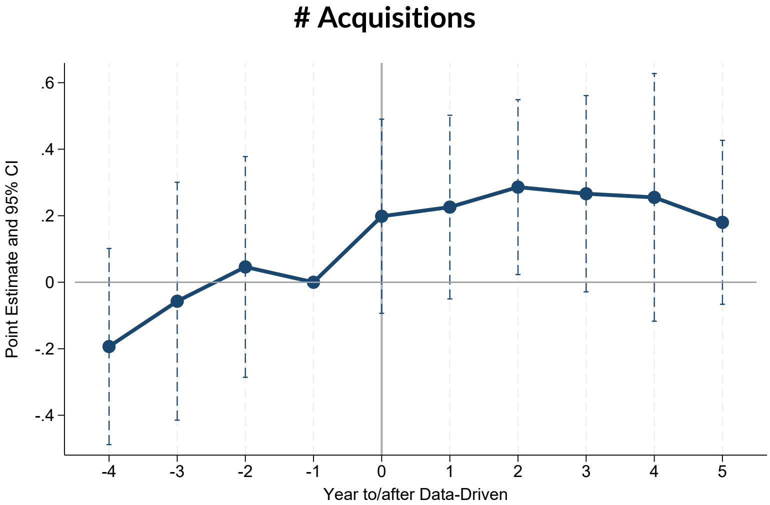
$$DataDriven_{j,t} = \beta VCExposure_j \times NewFund_{j,[-2:0]} + X_{j,t} + \alpha_j + \alpha_{c\times i\times s\times t} + \epsilon_{j,t}, \quad (1)$$

IV Results

Outcomes:		All	Non Hub	Low Activity
	First Stage	2SLS	2SLS	2SLS
	(1)	(2)	(3)	(4)
Data-Driven		0.696***	0.872***	1.436***
		(2.61)	(2.45)	(2.55)
VC Exposure × New Fund	0.055***			
	(3.71)			
Controls	Yes	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes	Yes
State × Industry × Stage × Year FE	Yes	Yes	Yes	Yes
F-Statistic	13.74			
R-squared		-0.04	-0.05	-0.02
N	3301	3301	3301	3301

back

Data Technologies & Post Investment Value Add



▸ back

Follow On Financing

Outcomes:	Follow On		
	(1)	(2)	(3)
Data Driven	0.026***	0.027***	0.029***
	(3.40)	(3.34)	(3.75)
Data Driven $\times$ Non Hub		-0.004	
		(-0.40)	
Data Driven $\times$ Low Activity			-0.163***
			(-3.61)
Non Hub		-0.002	
		(-0.27)	
Low Activity			-0.017
			(-1.05)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State $\times$ Industry $\times$ Stage $\times$ Year FE	Yes	Yes	Yes
Data-Driven = Data-Driven $\times$ Non Hub (p-value)		0.0405**	
Non Hub = Data-Driven $\times$ Non Hub (p-value)		0.8487	
Data-Driven = Data-Driven $\times$ Low Activity (p-value)			0.0001***
Low Activity = Data-Driven $\times$ Low Activity (p-value)			0.0023***
$\bar{y}$	0.61	0.61	0.61
R-squared	0.09	0.09	0.09
N	46871	46871	46871

back

IPO or Unicorn Status

Outcomes:	IPO or Unicorn Status		
	(1)	(2)	(3)
Data Driven	0.013	0.007	0.012
	(0.90)	(0.49)	(0.83)
Data Driven $\times$ Non Hub		0.026**	
		(2.25)	
Data Driven $\times$ Low Activity			0.045*
			(1.87)
Non Hub		-0.019*	
		(-1.79)	
Low Activity			-0.054***
			(-2.85)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State $\times$ Industry $\times$ Stage $\times$ Year FE	Yes	Yes	Yes
Data-Driven = Data-Driven $\times$ Non Hub (p-value)		0.0245**	
Non Hub = Data-Driven $\times$ Non Hub (p-value)		0.0077***	
Data-Driven = Data-Driven $\times$ Low Activity (p-value)			0.0058***
Low Activity = Data-Driven $\times$ Low Activity (p-value)			0.1408
$\bar{y}$	0.12	0.12	0.12
R-squared	0.16	0.16	0.16
N	22428	22428	22428

back

Acquisition

Outcomes:	Acquisition		
	(1)	(2)	(3)
Data Driven	0.002	0.007	0.002
	(0.09)	(0.29)	(0.08)
Data Driven $\times$ Non Hub		-0.022	
		(-0.98)	
Data Driven $\times$ Low Activity			0.013
			(0.14)
Non Hub		-0.018	
		(-1.30)	
Low Activity			-0.026
			(-0.89)
Controls	Yes	Yes	Yes
VC-Firm FE	Yes	Yes	Yes
State $\times$ Industry $\times$ Stage $\times$ Year FE	Yes	Yes	Yes
Data-Driven $=$ Data-Driven $\times$ Non Hub (p-value)		0.4333	
Non Hub $=$ Data-Driven $\times$ Non Hub (p-value)		0.8917	
Data-Driven $=$ Data-Driven $\times$ Low Activity (p-value)			0.9141
Low Activity $=$ Data-Driven $\times$ Low Activity (p-value)			0.7090
$\bar{y}$	0.23	023	0.23
R-squared	0.14	0.14	0.14
N	22428	22428	22428

back

Which non-hub areas attract data-driven investments?

- **Recap:** Data technology adoption $\downarrow$ search frictions; $\uparrow$ investments in non-hubs
- Advantage of algorithmic techniques: identify emerging trends and markets
  - More likely to invest in areas where there is more "data"
- Use the **Regional Entrepreneurship Cohort Potential Index** (*RECPI*)
  - $RECPI = SFR \times EQI$
    - SFR = Startup Formation Rate $\rightarrow$ quantity of new business registrants in an area
    - EQI = Entrepreneurship Quality Index $\rightarrow$ average growth potential within a group of startups
- ***Prediction:*** Commuting zones with high *RECPI* attract more data-driven investments

back

Commuting-Zone Level Investments

$$\underbrace{y_{c,t}}_{\text{\# investments}} = \beta \underbrace{Log(RECPI)_{c,t-1}}_{\substack{\text{entrepreneurial activity} \\ \text{statistic}}} + \underbrace{X_{c,t-1}}_{\text{cz controls}} + \underbrace{\alpha_c}_{\text{cz FE}} + \underbrace{\alpha_t}_{\text{year FE}} + \epsilon_{c,t}$$

Outcomes:	# Data Driven		
	Non Hub	Low Activity	
	(1)	(2)	
Log(RECPI)	0.815**	1.691*	
	(4.20)	(1.78)	
Controls	Yes	Yes	
Comzone FE	Yes	Yes	
Year FE	Yes	Yes	
R-squared	0.76	0.79	
N	5331	4598	

back

Commuting-Zone Level Investments

$$\underbrace{y_{c,t}}_{\text{\# investments}} = \beta \underbrace{Log(RECPI)_{c,t-1}}_{\text{entrepreneurial activity statistic}} + \underbrace{X_{c,t-1}}_{\text{cz controls}} + \underbrace{\alpha_c}_{\text{cz FE}} + \underbrace{\alpha_t}_{\text{year FE}} + \epsilon_{c,t}$$

Outcomes:	# Data Driven		
	Non Hub	Low Activity	
	(1)	(2)	
Log(RECPI)	0.815**	1.691*	
	(4.20)	(1.78)	
Controls	Yes	Yes	
Comzone FE	Yes	Yes	
Year FE	Yes	Yes	
R-squared	0.76	0.79	
N	5331	4598	

▸ back

Commuting-Zone Level Investments

$$\underbrace{y_{c,t}}_{\text{\# investments}} = \beta \underbrace{Log(RECPI)_{c,t-1}}_{\text{entrepreneurial activity statistic}} + \underbrace{X_{c,t-1}}_{\text{cz controls}} + \underbrace{\alpha_c}_{\text{cz FE}} + \underbrace{\alpha_t}_{\text{year FE}} + \epsilon_{c,t}$$

Outcomes:	# Data Driven		# Non-Data Driven	
	Non Hub	Low Activity	Non Hub	Low Activity
	(1)	(2)	(3)	(4)
Log(RECPI)	0.815**	1.691*	0.006	0.124
	(4.20)	(1.78)	(0.53)	(0.40)
Controls	Yes	Yes	Yes	Yes
Comzone FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
R-squared	0.76	0.79	0.91	0.61
N	5331	4598	5331	4598

back