6-1-2021

The Role of Data for AI Startup Growth

James Bessen

Stephen Michael Impink

Lydia Reichensperger

Robert Seamans

The Role of Data for AI Startup Growth

James Bessen, Technology & Policy Research Initiative, Boston University

Stephen Michael Impink, Stern School of Business, New York University

Lydia Reichensperger, Technology & Policy Research Initiative, Boston University

Robert Seamans, Stern School of Business, New York University

June 1, 2021

Revise & Resubmit for Research Policy, Special Issue on the Governance of AI

Abstract: Artificial intelligence ("AI")-enabled products are expected to drive economic growth. Training data are important for firms developing AI-enabled products; without training data, firms cannot develop or refine their algorithms. This is particularly the case for AI startups developing new algorithms and products. However, there is no consensus in the literature on which aspects of training data are most important. Using unique survey data of AI startups, we find that startups with access to proprietary training data are more likely to acquire venture capital funding.

JEL codes: O33, J21, L10, L26

Keywords: artificial intelligence, competition, data, algorithms, venture capital

1. Introduction

As described in the AI Index 2018 Annual Report (Shoham et al. 2018), artificial intelligence ("AI") has advanced rapidly over the past decade. Many scholars believe that AI has the potential to boost human productivity and economic growth (Athey 2018, Brynjolfsson et al. 2017, Furman & Seamans 2019). However, for this macroeconomic growth to be realized, firms pursuing AI products must gain access to the inputs needed to develop their products. The need for training data to train the algorithms underlying a firm's AI is important for all AI-producing firms. However, training data is especially important for AI startups. These startups need training data to develop effective AI products and scale. Without such data, these startups will have difficulties launching their initial product, raising venture capital ("VC") funds[1], and scaling their business model.

Data is not homogeneous; it can come from different sources and be used for different purposes. The choice of which training data to use is competitively significant as certain attributes of training data are a better fit with specific algorithms and products than others (Athey 2018, Bajari et al. 2018, Chiou & Tucker 2017, Donnelly et al. 2019, Varian 2014). The research question we ask in this paper is: what type of data is most important for AI startup growth? To address this question, we first argue that proprietary data—data that a firm can exclude others from using—is the most important type of data for AI startup growth. We then use responses to a recent survey to show that AI startup firms that use proprietary data receive more venture capital (VC) funding.

Our research makes several contributions. First, we contribute to a nascent stream of research on the role that data plays for AI-enabled firms (Bessen et al. 2018, Brynjolfsson et al. 2017, Cowgill et al. 2020, Furman & Seamans 2019) and a broader literature on digitization (Cowgill & Tucker 2019, Goldfarb & Tucker 2019, Jin & McElheran 2019, Jin et al. 2018, Savona 2019, Tucker 2019) by showing that the use of proprietary training data is related to future VC funding. We also find that the relationship between

[1] VC funding is an important determinant of startup performance (Kerr & Nanda 2009, Nanda 2016)

proprietary data and VC funding is stronger when firms are in markets where data provide a greater advantage. Lastly, our research provides practical information useful to policy-makers, designing policies to maximize innovation around AI while reducing potential negative externalities to consumers (e.g., ACCC 2019, Crémer et al. 2019, Furman et al. 2019), and managers, attempting to scale their AI startups.

This paper proceeds as follows. In the next section, we provide additional background from the academic literature on the connection between training data and competition, focusing on how training data as an input in production differs from other forms of information. Then, drawing on this literature, we provide our hypothesis that startups with access to proprietary data raise more VC funds in the future (Section 2). Next, we describe the data collected from surveying AI startups and corresponding measures (Section 3) and provide details on our research design (Section 4). We share our findings (Section 5), and then conclude (Section 6) by discussing limitations to our findings and methodology and describing the broader implications of this research.

2. Data as a Competitive Advantage

Scholars have studied more basic forms of the "sciences of the artificial" since the 1950s (Newell et al. 1954, Simon 1968, Turing 1950), developing mathematical models that enabled humans to see patterns in data. The advent of 'big data' and more sophisticated machine learning algorithms has brought AI products and needed training data to the forefront of many conversations around bias (Cowgill & Tucker 2019, Tucker 2019), fairness (Barocas et al. 2018, Mitchell et al. 2021), competition (Acquisti et al. 2016, Khan 2017, McSweeny & O'Dea 2018, Scott-Morton et al. 2019), and macroeconomic progress (Jones & Tonetti 2019, Farboodi et al. 2019). Despite this, we know little about how data as an input in production differs from other types of information. Recently, certain macroeconomic growth models started making a distinction between ideas (a set of instructions or processes) and data (all remaining forms of information) in efforts to understand the digital economy better (Jones & Tonetti 2019, Farboodi et al. 2019). However, these and other models still fail to account for the value of different types of data and various political economy implications (i.e., taxation, ownership rights) stemming from the concentration of data on a few large digital platforms (Savona 2020).

In the case of AI-enabled products, firms use their initial datasets to train their algorithms. These algorithms produce data as output which becomes an input in the next iteration in the algorithm training process, potentially exacerbating the importance of data inputs on competitive outcomes. As such, competitive analyses focused on end-products may undervalue data. Moreover, there is substantial debate about what data trains AI products better; there is no 'one size fits all' approach, as certain data may be better than others at training certain AI (Athey 2018, Donnely et al. 2019). Scholars argue that there are tradeoffs among quantity, quality, breadth, and recency of training data (Bajari et al. 2018, Varian 2014, Chiou & Tucker 2017). Training data spanning diverse groups are important to an algorithm's function, and there appear to be decreasing marginal returns to increased data quantity if those data are similar (Varian 2014, Bajari et al. 2018). Additionally, more recent data (i.e., the shorter lag time from collection to use) are particularly important in product or context search (Chiou & Tucker 2017).

Unlike large technology firms, AI startups do not have user-based platforms or other business lines that enable them to collect large amounts of data. Furthermore, even if they did, they might not have the complementary assets necessary to benefit from this additional data (Brynjolfsson & Hitt, 2000). For instance, computing power and human capital are also important to AI production. For computing power, AI startups rely on IT assets, either developed internally or licensed from a cloud services provider, which is important for startup survival, growth, and performance (Jin & McElheran 2019, Jin et al. 2018). Even high-potential startups may have difficulties paying for and developing physical IT assets internally (Nanda 2016, Nanda et al. 2020), potentially reducing the benefits from data to the firm (Farboodi et al. 2019). Additionally, AI startups rely on the market for human capital expertise; however, the largest firms have already established much cross-disciplinary expertise, including the highly specialized economics and machine learning expertise needed to analyze causal relationships and develop experiments (Thomke 2003, Varian 2014, Athey & Luca 2019) [2].

Even if startups can create their initial AI product innovation, they may not be able to develop a competitive advantage. For instance, they may lack the additional data, computational ability, or expertise necessary to benefit from follow-on complementary innovations, because these complementary innovations often rely on more of the same production inputs as the initial innovations (Brynjolfsson et al. 2017). Moreover, startups may lack needed proprietary resources, like R&D[3] or specific training data. For instance, if firms relied only on publicly available data (e.g., large data sets released by governments), they would be unable to exclude others from accessing and using the data. Proprietary data, on the other hand, provides the firm with an exclusionary right, enabling them to prevent others from using the data as an input in their production. Ultimately, the use of proprietary training data may enable startups to develop AI products using inputs in production that are less substitutable, making their products harder to replicate, more unique and therefore potentially more valuable.

For AI products to work correctly, training data must sufficiently fit the underlying technology and various dimensions of data quantity, breadth, and recency (Bajari et al. 2018, Varian 2014, Chiou & Tucker 2017). For an AI product to lead to a competitive advantage, we must also consider if its production inputs are replicable. AI startups utilizing only public data, a substitutable input, may be unable to create more differentiated products. Competitors using similar algorithms could acquire similar non-proprietary training data and create similar products, limiting the originating firm's ability to appropriate rents (isolating mechanism, Rumelt 1984) from their products (Peteraf 1997, Teece 1986, Barney 1987). However, startups utilizing proprietary data develop their products on production inputs that are less imitable and imperfectly substitutable. As a result, these startups can create less elastic, more differentiated products, impacting

[2] Google conducted 6,000 experiments on its search engine in 2008 (Varian 2014), and Amazon and Facebook conduct about 10,000 experiments per year (Athey & Luca 2019).

[3] Microsoft AI & Research and Google AI, that look more like an academic department than a business, offering a wealth of targeted, research-based advice on areas of corporate focus.

competitive outcomes and enabling them to raise more funds than their rivals without access to proprietary data. As such, we hypothesize that AI startups exclusively using proprietary data in AI product development experience more VC funding growth than AI startups that rely on public data.

3. AI Survey Data

We conducted an online survey in Qualtrics to reach founders, CEOs, CTOs, or other similar executives at 2,517 AI startups. Respondents to our surveys came from several sampling frames, the largest of which was from Crunchbase. From Crunchbase, we identified firms associated with the keyword "artificial intelligence" that have received funding, are in operation, and have not yet experienced an IPO. In addition to Crunchbase, we received a contact list of AI startups from the Creative Destruction Lab, a startup incubator based in Toronto, and another contact list from Philipp Hartmann and Joachim Henkel (Hartmann & Henkel 2018). Additionally, O'Reilly Media ran a notice of the survey in its AI newsletter, providing a link to the online survey.

Over 15-months, we received responses from 325 AI startups. We estimate that about five percent of the firms that we reached out to are not addressable in our study as they are located in China or no longer appear to be in business (bounce back from email), leading to a 13 percent response rate overall. Because the response rate is relatively low, one might worry that our respondent sample is biased. To address any biases arising from our respondent sample being different from the population, we use a Heckman selection correction in our results, as described below.

Ultimately, 271 firms responded to the survey question on the proprietary nature of training data, and 159 of these firms also have funding information in Crunchbase. We dropped two observations in which the respondent indicated that their firm was not involved with AI production. We report descriptive statistics in Table 1. We compare firm size from Crunchbase and from the responses to our survey and report the chart in Table A1. Note that responses are very similar (slightly higher frequency than Crunchbase for the smallest firms (<11 employees) and slightly lower for the second size tier of firms (11-50 employees)), suggesting that self-reported survey responses accurately depict respondents.

To test our hypothesis, we create measures from our survey responses and paired firm-level Crunchbase data. Through the survey, we collected information on if AI startups use proprietary data and data from customers to train their AI. The majority of firms (56 percent) use (a) some proprietary training data, defined as firm-held data collected to develop their products. This measure does not include firms that only use proprietary data sourced from customers. Next, we create an additional dummy variable for respondents using (b) any customer data, defined as data sourced from their customers (79 percent). Since we do not know the exact nature of the data-sharing arrangements between the firm and its customers, customer data is a mix of proprietary data and public data sourced through customer telemetry or direct data agreements. Moreover, customer data could include data about a customer's customer. Third, we create dummy variables for firms using (c) a mix of firm-held proprietary and customer data (41 percent). For instance, some firms use customer data but do not respond that they use propriety data. Lastly, we have a dummy variable for firms that only use (d) proprietary data only without any customer data (15 percent), responding to only the first answer choice in the provided survey question in Appendix A2. We provide a summary of these measures in this appendix and descriptive statistics in Table 1.

We rely on funding data from Crunchbase to create a measure of VC funding after 2019 (i.e., after the survey concluded). Even though our models are cross-sectional, this timing variance enables us to examine how access to proprietary training data in an earlier period could impact funding performance in a later period. We also include a control for prior VC funding before 2019. Additionally, we collected and use firm age, firm age^2, firm size, and geographic locations (e.g., dummy variables for the United States, Canada, and Germany) from Crunchbase as additional controls in our models. We report these measures and their summary statistics in Table 1. We also report the correlation of these measures with firm demographics and performance measures in the Appendix in Table A3.A to A3.C.

4. Research Design

4.1. Selection. We use regression models to explore the relationship between proprietary training data and VC funding. We use Heckman's selection approach (Heckman 1976, 1979) and Coarsened Exact Matching (CEM, Iacus et al. 2019) to help address selection and endogeneity issues. First, given our lower survey response rate and reliance on cross-sectional data, we analyze if our survey respondents are similar to the broader population of startups in Crunchbase. From the t-test results, we find that responses from New York are over-represented, and responses from small firms (<10 employees) and California-based startups are under-represented, reported in Table 1. The probit model confirms that startups that are very small or are located in California are less likely to respond; startups located in New York are more likely to respond[4] (Table 2).

Based on this, we use Heckman's two-step procedure to account for selection issues from possible respondent missingness to support the argument that our sample of respondents does not bias our main OLS model estimates. We include dummy variables for small firm size and HQ locations in New York and California in the first step, below, to obtain estimates of γ.

$$\textit{(1)} \qquad response_i = w_i\gamma + \mu \qquad\qquad \text{[selection equation]}$$

where,

$response$ takes the value of 1 if a firm in the population responds to the survey, otherwise 0.

w_i is a vector of firm demographic dummy variables (e.g., NY, California, Small (<11 employees)) that are plausibly correlated with sample response.

Now that we have obtained the estimates of γ from the selection equation, we compute the inverse Mills ratios of each observation.

[4] We use all AI startups that we contacted (2,517 firms) as the population with a dummy variable for firms that responded and are observed (271 firms).

$$(2) \qquad \lambda = \frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)} \qquad \text{[inverse Mill's ratio]}$$

where,

$\phi(w_i\gamma)$ is the probability density function

$\Phi(w_i\gamma)$ is the complementary cumulative distribution function

Next, we use CEM to ensure that the firms using proprietary data are observationally similar to those not using proprietary data. We include HQ locations (dummies for the US and EU), age, and employment size as parameters in the CEM model and match 150 firms of the 159 firms with both survey and funding data. The match reduces the difference in standardized means across these observable demographic variables between the respondents who use and do not use proprietary data. We show the differences in standardized means for these groups before and after matching in the Appendix in Figure A4.

4.2. Main OLS Specification. For the main regression, we use OLS to provide linear approximations since we have a continuous dependent variable (Angrist & Pischke 2009, Gibson 2019). As controls in the main regression, we include firm age, which is often related to acquiring funding outcomes. Older firms have more opportunities to raise funds. Additionally, we use age^2 to adjust for the curvilinear relation that likely exists between age and funding. For instance, beyond a certain age, older firms that have not experienced significant growth may no longer be considered for VC investment (Kerr & Nanda 2009). We also include the log of prior funding to help control for the unobservable variables influencing prior funding that may also be correlated with increased future funding. Lastly, we include a dummy variable for having their headquarters in the San Francisco area, where most VC funds are located. The dummy variable for San Francisco is correlated (69 percent) with the dummy variable for California; however, this is the only control variable in the main OLS model that is highly correlated with a variable used in the Heckman first-stage. We do not include controls for firm size (<11 employees) or New York because they are used in the Heckman first-stage selection equation.

(3) $funding_i = \beta_0 + \beta_1 data_i + \lambda + \rho + \mu$ [mainregression equation]

where,

$funding_i$ is the log of the funding raised between January 1, 2019 and May 31, 2020

$data_i$ for takes the value 1 if a firm responds to using existing proprietary data, 0 otherwise

ρ are controls for age, age^2, prior VC funding before 2019, San Francisco HQ dummy variable

λ is the inverse of the Mills ratio, included controlling for representativeness of our sample compared with the population of AI startups that we sourced and contacted

μ is the error term

5. Findings

Our main results are reported in Table 3. These models rely on both Heckman's two-stage selection procedure to control for non-response and matching (CEM) to support that firms using firm proprietary data are observationally similar to those using non-proprietary data. In model (1), the base model, we investigate the relationship between proprietary data and VC funding. The coefficient on proprietary data is positive and statistically significant. In model (2) we include the inverse Mill's ratio from the first stage of Heckman's procedure. In model (3) we repeat the first model, including log VC funding prior to 2019 to control for aspects of prior performance that may be endogenous with funding outcomes. The coefficient on proprietary data remains positive and statistically significant.

We then run two more models with additional controls. In model (4) we include a control for firm age and firm age^2. The coefficient on age measures (age is positive and age^2 is negative) suggests a curvilinear effect of age on funding (i.e., funding increased with age until a certain point and then decreases). So, certain older firms may still exist as entrepreneurial ventures but are not high-potential startups. Lastly, in model (5) we include a dummy variable for startup location in the San Francisco bay area, where there is a high concentration of venture capital firms. In summary, we find that the use of proprietary data is related to increased future funding in all these models. This effect holds when using Heckman's selection, matching (CEM), and controls for prior performance and aspects of firm demography related to funding (age and location) (Table 3, model (5): +2.7 SD 1.0).

We then examine if firms using a mix of propriety data and customer data experience a similar positive relationship with increased future funding. To do this, we include two measures, (i) Proprietary and Customer Data and (ii) Only Proprietary Data (with no customer data). Using only proprietary data are related to a larger increase in future VC funding than using a mix of customer and proprietary data (Table 3, model (6): Prop & Customer +1.9 SD 1.1; Only Prop Data, +4.8 SD 1.7). To further support this, we run a separate model including only proprietary data without any customer data and again find a stronger positive correlation with increased future funding (+3.9 SD 1.6, Appendix, Table A5, model (5)). Proprietary data from customers do not provide the same advantage as other forms of firm-held proprietary data, suggesting that the source of the training data may be competitively significant.

Next, we examine mechanisms that may impact the relationship between proprietary data and increased funding and report these results in Table 4. First, in model (1) we include the main result from Table 3, model (5) as a point of comparison for the next several models. In a subset of firms (120 of the initial 150 matched firms), we have additional survey responses on if data ownership provides a major advantage in their markets.[5] In model (2) we replicate our main specification on the subset of 120 matched firms with additional survey data, which provides very similar results as in model (1).

In model (3) we include an interaction between firms using proprietary training data and firms that respond that owning data is a major advantage in their market. The coefficient on this interaction term is positive and significant (+4.7 SD 2.3, model (3)). We graph this interaction in the Appendix in Table A8. Next, in models (4) and (5) we present split sample results. We show that proprietary data is related to even

[5] This survey question is also included in Appendix, Note A2.

higher future funding amounts (+4.5, SD 1.9, model (4)) in markets where owning data is a major advantage. Alternately, for firms responding that owning data is not a major advantage in their markets, we find no significant effect (model (5)).

We conduct a variety of robustness checks and present these results in an Appendix for space considerations. We show that results are similar for the full sample of firms (159 firms) without any CEM matching (but still including the Heckman selection approach) in the Appendix in Table A6. In our main results, the inverse Mill's ratio is not significant, but this does not necessarily mean that respondent selection bias does not exist (Certo et al. 2016). Given this, we include the inverse Mill's ratio as a control in the main results, but also reproduce our findings without this control (i.e., without using Heckman's two-stage selection procedure) on the matched sample (150 firms) in the Appendix in Table A7. In these tables, we use a similar buildup of models adding control variables (base model, prior funding, age, and location in San Francisco) as we did with the main results. The results are consistent with the results in the main text.

Next, we examine the same specification with an alternative dependent variable, creating a dummy variable for increased VC funding, if the funding CAGR is higher in the post-period, 2019 and after, than in the earlier period, before 2019. We find a positive, significant effect in the OLS specification (Table A9, model (3): 0.23 SD .06). Since the dependent variable is binary, we also run a probit regression that also supports a positive, significant relationship (Table 9, model (6): +0.95 SD 0.3). The results are consistent with our main results.

6. Conclusion

In this study, we find a significant positive correlation between proprietary training data and future VC funding. We also find that firms benefit more from proprietary data when they are in markets where owning data provides an advantage. Additionally, data from customers may not provide the same level of benefit as proprietary data from other sources. We believe these empirical findings support the idea that proprietary data are imperfectly substitutable inputs in production. As such, using proprietary training data leads to less imitable products, positively impacting a startup's ability to collect additional rents from the market and develop an initial competitive advantage in this nascent industry.

Our results are derived from cross-sectional survey data, which has its limitations. We have attempted to address these issues as much as possible by using Heckman selection correction and Coarsened Exact Matching approaches. Though we control for prior fundraising, age, location near San Francisco, where there is a high concentration of VC firms, we cannot entirely rule out that well-funded firms can access proprietary data more easily or that we are not capturing other unobservable aspects of performance correlated with future funding. For instance, one possibility is that our measures capture the ability of a startup to forge partnerships with other firms, such as large cloud services providers. Another possibility is that our measures endogenously capture some elements of founder connections or leadership ability. Additionally, there could be other unobservable relationships, such as a relationship with higher status or reputation venture capital or the impact of accelerator programs that assist in proprietary data acquisition. For instance, serial entrepreneurs could have proprietary data from another venture, or groups of startups could pool data resources. These are all issues that future research could investigate.

Regulating data remains a topic of intense debate, especially amid recent policies that limit data availability to protect consumer privacy, including the European Union’s General Data Protection Regulation (“GDPR”) and California’s Consumer Privacy Act (“CCPA”). In many cases, there is a tradeoff between increased data regulation and access to training data, asymmetrically increasing the costs of collecting and using data for smaller firms (Bessen et al. 2020b, Johnson et al. 2020). Regulation may specify that certain data are deleted or withheld from use, increasing the scarcity and value of training data for AI startups. Many governments recognize the importance of AI advancements to the broader economy and may consider establishing policies to increase the entry of AI startups, such as data sharing (Calo, 2017, Himel and Seamans, 2017). However, regulations that require data sharing may not necessarily increase entry if the training data are fully substitutable. Firms may hesitate to invest in developing products with training data that is not proprietary, and VCs may hesitate to invest in firms that don’t provide innovative or differentiated products. Our findings suggest that proprietary data, not public data widely available to many firms, is what drives follow-on VC funding. Thus, some skepticism is warranted that policies around data sharing will help drive AI startup success.

Table 1 - Summary Statistics

	Survey Respondents (177 obs)				Sample CB (2,517 obs)				T-Test (p)
	Mean	SD	Min	Max	Mean	SD	Min	Max	Resp/CB
Survey Question #15									
Any Proprietary Data	0.56	0.50	0.00	1.00					
Any Customer Data	0.79	0.41	0.00	1.00					
Proprietary Data (No Customer Data)	0.15	0.38	0.00	1.00					
Proprietary and Customer Data	0.41	0.49	0.00	1.00					
Demographics									
Size & Age									
Employee Size (buckets, 4)	1.78	0.71	1.00	4.00	1.73	0.79	1.00	4.00	0.41
Employee Size (dummy, small <11)	0.36	0.48	0.00	1.00	0.44	0.50	0.00	1.00	0.04
Employee Size (dummy, large >100)	0.12	0.32	0.00	1.00	0.12	0.32	0.00	1.00	0.94
Firm Age (cont.)	4.35	2.05	0.17	10.00	4.44	1.91	0.16	10.00	0.55
Firm Age (dummy, <2 Years)	0.16	0.37	0.00	1.00	0.13	0.33	0.00	1.00	0.19
Firm Age (dummy, >5 Years)	0.40	0.49	0.00	1.00	0.37	0.48	0.00	1.00	0.55
Geography									
Region (buckets, 5)	2.77	1.21	1.00	5.00	3.01	1.16	0.00	1.00	0.01
US (dummy)	0.38	0.49	0.00	1.00	0.45	0.50	0.00	1.00	0.06
California (dummy)	0.33	0.47	0.00	1.00	0.24	0.43	0.00	1.00	0.01
San Francisco (dummy)	0.09	0.29	0.00	1.00	0.13	0.34	0.00	1.00	0.05
Mass (dummy)	0.02	0.15	0.00	1.00	0.03	0.16	0.00	1.00	0.79
NY (dummy)	0.03	0.18	0.00	1.00	0.07	0.26	0.00	1.00	0.01
Germany (dummy)	0.03	0.17	0.00	1.00	0.04	0.19	0.00	1.00	0.58
Canada (dummy)	0.04	0.20	0.00	1.00	0.05	0.22	0.00	1.00	0.49
Funding									
Funding 2019 & After	4.33	6.67	0.00	17.90	5.51	7.32	0.00	20.21	0.04
Funding Before 2019	9.57	6.70	0.00	17.50	7.86	7.10	0.00	18.89	0.04
Funding Growth Increase (2019)	0.22	0.12	0.00	1.00	0.30	0.46	0.00	1.00	0.02

Notes: On Survey Measures, Any Proprietary Data (response 1, yes), Any Customer Data (response 2, yes, or response 3, yes), Proprietary Data (No Cust) (response 1 yes and response 2, no, & reponse 3, no), Proprietary Cust Data (response 1, yes, & (response 2, yes, or response 3, yes)). Five regions includes are North America, South America, Europe, Middle East & Africa, and Asia.

Table 2 - Base Probit for Heckman

	(1)	(2)	(3)
		Response (dummy)	
Employ. Small (<11)	-0.155**	-0.155**	-0.148*
	(0.079)	(0.079)	(0.079)
California		-0.368*	-0.311
		(0.188)	(0.190)
NY			0.196**
			(0.086)
Observations	2331	2331	2331
Psuedo R2	0.03	0.06	0.11

Notes: * p<0.1, ** p<0.05, *** p<0.01. Coefficients are estimated using Probit regression, showing the buildup to model (3), which supports the variables used in the first stage of the Heckman selection procedure.

Table 3 - Proprietary Data & Funding (Heckman & CEM)

	(1)	(2)	(3)	(4)	(5)	(6)
DV is log of:	**Funding 2019 & After**					
Any Prop Data	2.479**	2.477**	2.225**	2.396**	2.657**	
	(1.061)	(1.063)	(1.031)	(1.019)	(1.023)	
Prop & Customer Data						1.905*
						(1.101)
Only Prop Data						4.821***
						(1.679)
IMR		1.031	-1.070	-2.752	0.166	0.339
		(4.768)	(5.003)	(5.117)	(5.356)	(5.245)
log(Funding before 2019)			-0.249***	-0.171*	-0.194*	-0.173*
			(0.095)	(0.101)	(0.100)	(0.097)
log(Age)				29.357***	28.468***	37.580***
				(10.270)	(10.485)	(11.019)
log(Age^2)				-13.801***	-13.305***	-17.277***
				(4.866)	(4.975)	(5.110)
SF					3.519	3.513
					(2.271)	(2.254)
Observations	150	150	150	150	150	150
Adj. R2	0.0291	0.0228	0.0763	0.0914	0.105	0.121

Notes: * p<0.1, ** p<0.05, *** p<0.01. Coefficients are estimated using OLS regression and include robust standard errors, in parentheses below the coefficient. All models include matching (CEM), based on firm age (cont.), employment size (buckets, employment small), and region (US, EU), dropping 9 firms. Additionally, all models use Heckman's selection procedure, controlling with IMR.

Table 4 – Proprietary Training Data and "Major Advantage from Data"

	(1)	(2)	(3)	(4)	(5)
DV is log of:	**Funding 2019 & After**				
Sample	All	Survey 2 only	Survey 2 only	Survey 2 & Data Maj. Adv.	Survey 2 & Data Not Maj. Adv.
Any Prop Data	2.657**	2.328*	-0.426	4.527**	-0.337
	(1.023)	(1.206)	(1.582)	(1.930)	(1.464)
IMR	0.166	1.767	2.686	13.614	-8.952
	(5.356)	(5.735)	(5.549)	(8.762)	(5.729)
log(Funding before 2019)	-0.194*	-0.173	-0.131	-0.024	-0.309**
	(0.100)	(0.105)	(0.102)	(0.165)	(0.126)
log(Age)	28.468***	28.300***	36.373***	40.601**	145.136*
	(10.485)	(10.659)	(10.521)	(15.807)	(82.802)
log(Age^2)	-13.305***	-13.345***	-17.006***	-19.461**	-62.546*
	(4.975)	(5.092)	(4.995)	(7.642)	(35.555)
SF	3.519	2.931	3.008	3.085	1.984
	(2.271)	(2.436)	(2.260)	(2.840)	(3.848)
Data Major Adv.			-0.776		
			(1.624)		
Any Prop x Data Major Adv.			4.720**		
			(2.315)		
Observations	150	120	120	61	59
Adj. R2	0.105	0.0753	0.107	0.157	0.0854

Notes: * p<0.1, ** p<0.05, *** p<0.01. Coefficients are estimated using OLS regression, and include robust standard errors, in parentheses below the coefficient. All models include CEM, based on firm age (cont.), employment size (buckets) and regions (buckets), dropping 9 firms. Additionally, all models use Heckman's selection procedure, controlling with IMR.

References

Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*, 54(2), 442-92.

Angrist, J. D., J. Pischke. (2009) Mostly harmless econometrics: An Empiricist's Companion. Princeton University Press.

Athey, S. (2018). The Impact of Machine Learning on Economics. In the Economics of Artificial Intelligence: An Agenda. *University of Chicago Press.* (507-547)

Athey, S., & Luca, M. (2019). Economists (and Economics) in Tech Companies. *Journal of Economic Perspectives*, 33(1), 209-30.

Australian Competition and Consumer Commission (ACCC): Digital Platforms Inquiry. Final Report, June 2019.

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2018). The Impact of Big Data on Firm Performance: An Empirical Investigation. *NBER Working Paper* (No. w24334).

Barney, J. (1991). Special Theory Forum the Resource-based Model of the Firm: Origins, Implications, and Prospects. *Journal of Management,* 17(1), 97-98.

Bessen, J. E. (2016). How Computer Automation Affects Occupations: Technology, Jobs, and Skills. *Boston University School of Law, Law and Economics Research Paper*, (15-49).

Bessen, J. E., Impink, S. M., Reichensperger, L., & Seamans, R. (2018). The Business of AI Startups. *Boston University School of Law, Law and Economics Research Paper*, (18-28).

Bessen, J. E., Impink, S. M., Reichensperger, L., & Seamans, R. (2020b). GDPR and the Importance of Data to AI Startups. *SSRN Working Paper*. 3576714.

Bessen, J., Goos, M., Salomons, A. and van den Berge, W. (2020a). Firm-Level Automation: Evidence from the Netherlands. *American Economic Association Paper and Proceedings*, May 2020.

Brynjolfsson, E., & Hitt, L. M. (2000). Beyond Computation: Information Technology, Organizational Transformation and Business Performance. *Journal of Economic Perspectives*, 14(4), 23-48.

Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. *NBER Working Paper* (No. w24001).

Brynjolfsson, E., Rock, D., & Syverson, C. (2019). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *SSRN Working Paper.* 3346739.

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. UCDL Rev., 51, 399.

Certo, S. T., Busenbark, J. R., Woo, H. S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, 37(13), 2639-2657.

Chiou, L., & Tucker, C. (2017). Content Aggregation by Platforms: The Case of the News Media. *Journal of Economics & Management Strategy*, 26(4), 782-805.

Cowgill, B., & Tucker, C. E. (2019). Economics, Fairness and Algorithmic Bias. Preparation for*: Journal of Economic Perspectives*.

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020). Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation* (679-681).

Crémer, J., de Montjoye, Y. A., & Schweitzer, H. (2019). Competition Policy for the Digital Era. Report for the European Commission.

Donnelly, R., Ruiz, F. R., Blei, D., & Athey, S. (2019). Counterfactual Inference for Consumer Choice Across Many Product Categories. arXiv:1906.02635.

Farboodi, M., & Veldkamp, L. (2019). A Growth Model of the Data Economy. *Columbia Business School Working Paper*, New York, June 2020.

Furman, J., & Seamans, R. (2019). AI and the Economy. *Innovation Policy and the Economy*, 19(1), 161-191.

Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019). Unlocking digital competition: Report of the Digital Competition Expert Panel. *UK Government Publication, HM Treasury*.

Gibson, J. (2019). Are You Estimating the Right Thing? An Editor Reflects. *Applied Economic Perspectives and Policy*, 41(3), 329-350.

Goldfarb, A., & Tucker, C. (2019). Digital Economics. *Journal of Economic Literature*, 57(1), 3-43.

Hartmann, P., & Henkel, J. (2018). Really the New Oil? A Resource-based Perspective on Data-driven Innovation. *Academy of Management Global Proceedings*, (2018), 142.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, v. 5, n. 4 (pp. 475-492).

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 153-161.

Himel, S., & Seamans, R. (2017). Artificial Intelligence, Incentives to Innovate, and Competition Policy. Antitrust Chronicle, 1(3).

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 1-24.

Jin, W., & McElheran, K. (2019). Economies Before Scale: Survival and Performance of Young Plants in the Age of Cloud Computing. *Rotman School of Management Working Paper* (3112901).

Johnson, G. A., Shriver, S. K., & Du, S. (2020). Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry?. *Marketing Science*.

Jones, C. I., & Tonetti, C. (2019). Nonrivalry and the Economics of Data. *NBER Working Paper* (No. w26260).

Kerr, W., & Nanda, R. (2009). Financing constraints and entrepreneurship. *NBER Working Paper* (No. w15498).

Khan, L. M. (2016). Amazon’s Antitrust Paradox. *Yale Law Journal*, 126, 710.

McSweeny, T., O'Dea, B. (2018). Data, Innovation, and Potential Competition in Digital Markets. *CPI Antitrust Chronicle*, February 2018.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint.* arXiv:1811.07867.

Nanda, R. (2016). Financing High-Potential Entrepreneurship. *IZA World of Labor*. 05530.

Nanda, R., Samila, S., & Sorenson, O. (2020). The persistent effect of initial success: Evidence from venture capital. *Journal of Financial Economics*.

Newell, A., Simon, H. A., & Shaw, J. C. (1954). Papers on Artificial Intelligence and Cognitive Processes. *Rand Corporation*.

Peteraf, M. A. (1993). The Cornerstones of Competitive Advantage: A Resource-based View. *Strategic Management Journal*, 14(3), 179-191.

Rumelt, R. P. (1984). Towards a Strategic Theory of the Firm. *Competitive Strategic Management*, 26(3), 556-570.

Savona, M (2019), The Value of Data: Towards a Framework to Redistribute It. *SPRU Working Paper* 2019-21.

Savona, M. (2020). Governance Models for Redistribution of Data Value. *VOX, CEPR Policy Portal*, 17 January 2020

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., ... & Bauer, Z. (2018). *The AI Index 2018 Annual Report*. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA.

Simon, H. (1968). Mathematical Models and Artificial Intelligence. *Brain Function and Learning*, 4, 169-209.

Teece, D. J. (1986). Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy. *Research Policy*, 15(6), 285-305.

Thomke, S. H. (2003). Experimentation Matters: Unlocking the Potential of New Technologies for Innovation. *Harvard Business Press*.

Tucker, C. (2019). Digital Data, Platforms and the Usual [Antitrust] Suspects: Network effects, Switching Costs, Essential facility. *Review of Industrial Organization*, 54(4), 683-694.

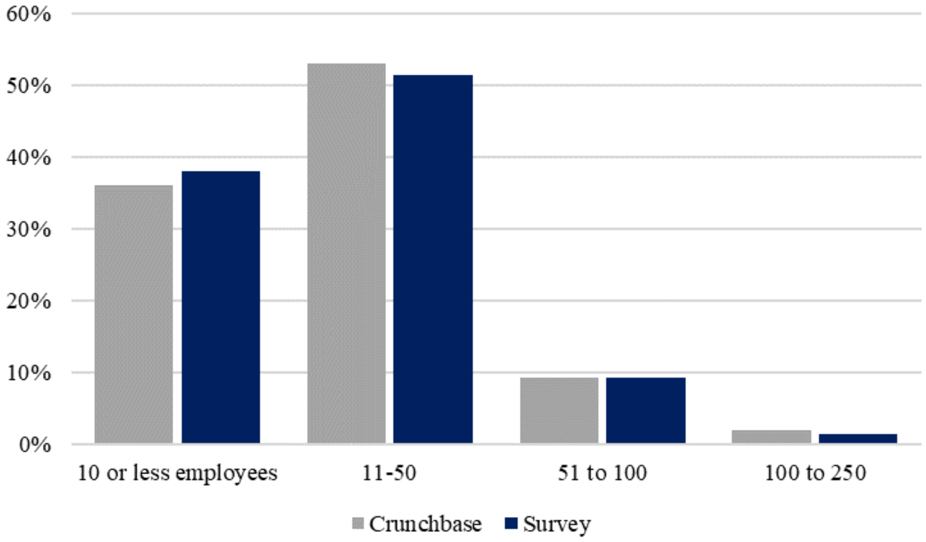
Turing, A. (1937) On Computable Numbers, with Applications to the Entscheidungs Problem. *London Math Society Proceedings*, (2) 42 (1937), 230-265.

Turing, A. M. (1950). Can a Machine Think?. *Mind*, 59(236), 433-460.

Varian, H. R. (2014). Beyond Big Data. *Business Economics*, 49(1), 27-31.

Appendix

Figure A1 - Firm Size Self-reported in Survey vs. Crunchbase



Note A2 - Survey Questions and Measure Creation

Proprietary Data Question

Which of the following types of data does your product rely on?
(Please select all that apply.)

□ Your firm's proprietary data (1)

□ Your customer's data about their customers and users (2)

□ Other proprietary data from your customer (3)

□ Other third-party data provider (4)

□ Publicly available data (including demographic data from government agencies or data scraped from the internet) (5)

□ Publicly available benchmarks for artificial intelligence (e.g., CIFAR) (6)

□ Synthetic data (8)

□ No data needed (7)

Additional Details on Measure Creation

(a) Any Proprietary Data. Firms using any firm-held proprietary training data to train their AI (56 percent) [checks box 1]

(b) Any Customer Data. Firms that use some data sourced from their customers (79 percent) [checks box 2, 3 or both box 2 and 3]

(c) Proprietary and Customer Data. Firms using a mix of firm-held proprietary and customer data (41 percent) [checks box 1 and either box 2, 3, or both box 2 and 3]

(d) Only Proprietary Data (No Cust.). Firms using only firm-held proprietary training data without any customer data (15 percent) [check only box 1]

Data Advantage Question

How strong of an advantage does ownership of data provide in your market?

○ No advantage (1)

○ Minor advantage (3)

○ Major advantage (4)

○ I don't know (99)

A3. Correlation Tables

Table A3.A - Correlation Table (Funding)

	Any Prop Data	Any Customer Data	Prop (No Cust)	Prop & Customer	After 2019 Funding	Before 2019 Funding
Any Customer Data	-0.2028*					
	0.0069					
Only Prop Data (No Cust)	0.3844*	-0.8315*				
	0.0000	0.0000				
Prop & Customer	0.7496*	0.3883*	-0.3229*			
	0.0000	0.0000	0.0000			
Funding 2019 & After	0.1600*	-0.1566*	0.2014*	0.0154		
	0.0432	0.048	0.0106	0.8465		
Before 2019 Funding	0.0056	0.0115	-0.0653	0.0532	-0.1245*	
	0.9436	0.8849	0.4119	0.5042	0.0016	
Funding Growth Increase (2019, dummy)	0.2292*	-0.1039	0.1588*	0.1164	0.8702*	-0.0498
	0.0036	0.191	0.0449	0.1427	0.0000	0.2081

Notes: * p<0.1

Table A3.B - Correlation Table (Age & Size)

	Any Prop Data	Any Customer Data	Prop (No Cust)	Prop & Customer	Employ. Size	Employ. Small
Any Customer Data	-0.2028*					
	0.0069					
Only Prop Data (No Cust.)	0.3844*	-0.8315*				
	0.0000	0.0000				
Prop & Customer	0.7496*	0.3883*	-0.3229*			
	0.0000	0.0000	0.0000			
Employ. Size (4)	0.074	0.111	-0.1025	0.1493*		
	0.3289	0.1424	0.176	0.0479		
Employ. Small (<11)	-0.114	-0.1686*	0.1323	-0.2117*	-0.8099*	
	0.1319	0.0253	0.0801	0.0048	0.0000	
Age (cont.)	0.0314	0.0157	-0.0181	0.0452	0.3004*	-0.2795*
	0.6789	0.8362	0.8118	0.5516	0.0000	0.0000

Notes: * p<0.1

Table A3.C - Correlation Table (Geography)

	Any Prop Data	Any Customer Data	Prop (No Cust)	Prop & Customer	Region	US	Calif.	SF
Any Customer Data	-0.2028*							
	0.0069							
Only Prop Data (No Cust.)	0.3844*	-0.8315*						
	0.0000	0.0000						
Prop & Customer	0.7496*	0.3883*	-0.3229*					
	0.0000	0.0000	0.0000					
Regions (5)	0.1076	0.0054	-0.0096	0.117				
	0.1565	0.9437	0.8994	0.1229				
US	0.1164	-0.0649	0.0529	0.0811	0.7836*			
	0.1251	0.3937	0.4865	0.2858	0.0000			
California	0.1187	-0.1227	0.101	0.0493	0.4915*	0.6276*		
	0.1166	0.1048	0.1823	0.5158	0.0000	0.0000		
San Francisco	-0.0576	0.0046	-0.0721	-0.0074	0.3364*	0.4295*	0.6847*	
	0.4477	0.9522	0.3419	0.9226	0.0000	0.0000	0.0000	
New York	-0.0107	0.0919	-0.0764	0.0438	0.2350*	0.3000*	-0.1530*	-0.1048*
	0.888	0.225	0.3133	0.5634	0.0000	0.0000	0.0000	0.0000

Notes: * p<0.1

Figure A4 - CEM Matching Comparison, Before and After Matching

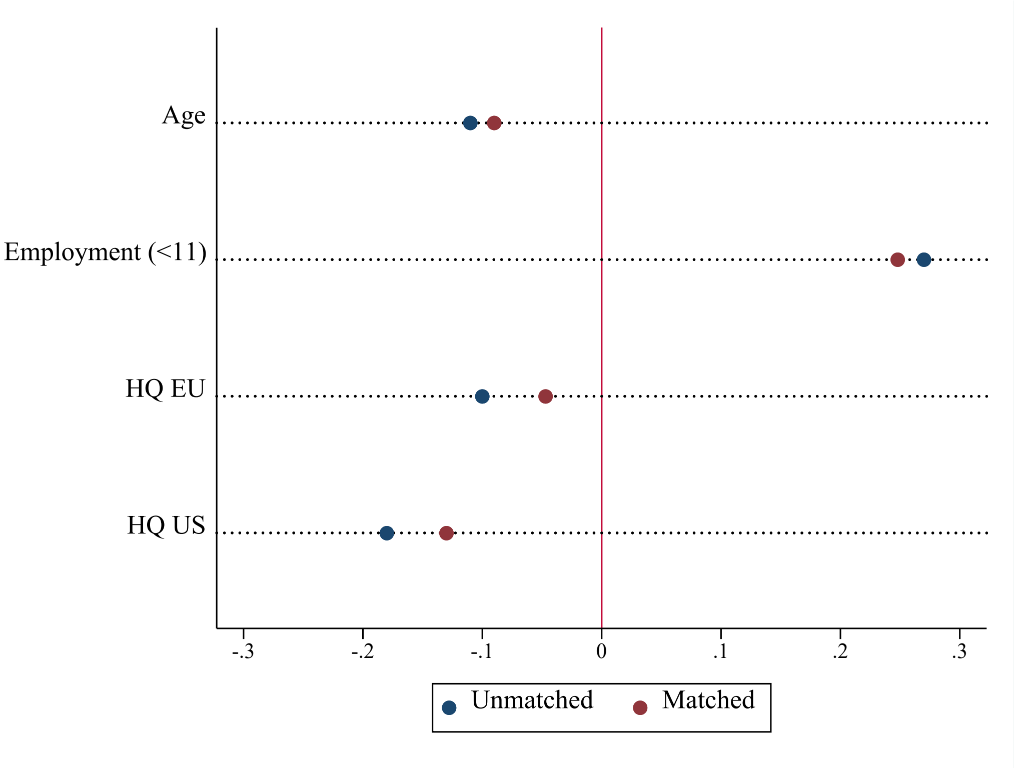


Table A5 - Prop Data (No Customer Data) & Funding (Heckman & CEM)

	(1)	(2)	(3)	(4)	(5)
DV is log of:			**Funding 2019 & After**		
Only Prop Data (No Cust.)	3.763**	3.769**	3.244*	3.763**	3.877**
	(1.683)	(1.695)	(1.642)	(1.645)	(1.620)
IMR		1.295	-0.763	-2.469	0.086
		(4.827)	(5.024)	(5.083)	(5.335)
log(Funding before 2019)			-0.240***	-0.156	-0.177*
			(0.092)	(0.095)	(0.094)
log(Age)				38.415***	37.619***
				(10.893)	(10.991)
log(Age^2)				-17.717***	-17.272***
				(5.049)	(5.103)
SF					3.063
					(2.376)
Observations	150	150	150	150	150
Adj. R2	0.0368	0.0307	0.0795	0.0996	0.109

Notes: * p<0.1, ** p<0.05, *** p<0.01. Coefficients are estimated using OLS regression, and include robust standard errors, in parentheses below the coefficient. All models include matching (CEM), based on firm age (cont.), employment size (buckets, employment small) and region (US, EU), dropping 9 firms. Additionally, all models use Heckman's selection procedure, controlling with IMR.

Table A6 - Proprietary Data & Funding (Heckman, No CEM)

	(1)	(2)	(3)	(4)	(5)	(6)
DV is log of:	**Funding 2019 & After**					
Any Prop Data	2.086**	2.005*	1.939*	2.150**	2.546***	
	(1.028)	(1.038)	(0.994)	(0.978)	(0.958)	
Prop & Customer Data						1.710
						(1.036)
Only Prop Data (No Cust.)						4.982***
						(1.588)
IMR		-2.437	-3.971	-5.544	-1.253	-1.089
		(4.684)	(4.757)	(4.799)	(5.001)	(4.885)
log(Funding before 2019)			-0.258***	-0.169*	-0.181**	-0.164*
			(0.084)	(0.093)	(0.089)	(0.086)
log(Age)				34.969***	33.147***	43.247***
				(10.150)	(10.747)	(11.132)
log(Age^2)				-16.389***	-15.489***	-19.866***
				(4.783)	(5.043)	(5.126)
SF					4.710**	4.806**
					(2.176)	(2.154)
Observations	159	159	159	159	159	159
Adj. R2	0.0182	0.0137	0.0747	0.100	0.127	0.149

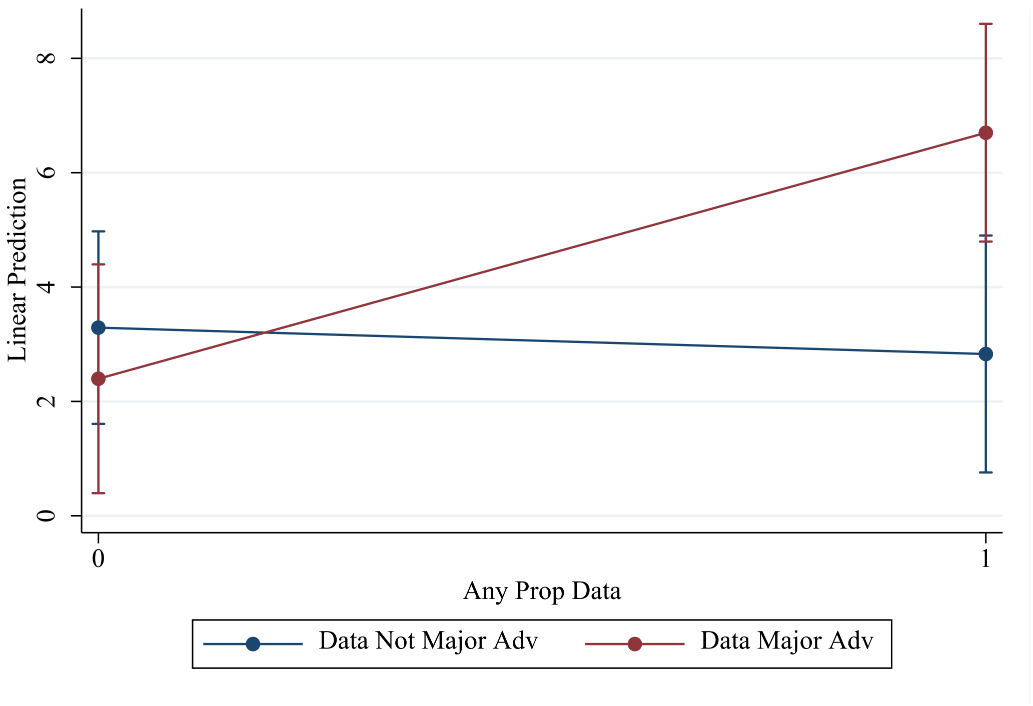
Notes: * p<0.1, ** p<0.05, *** p<0.01. Coefficients are estimated using OLS regression, and include robust standard errors, in parentheses below the coefficient. Models include IMR from the first-stage Heckman Selection, including employment small (<11), California, and New York (dummies). All these models do not include matching (CEM).

Table A7 - Proprietary Data & Funding (CEM, No Heckman)

	(1)	(2)	(3)	(4)	(5)
DV is log of:			**Funding 2019 & After**		
Any Prop Data	2.479**	2.226**	2.391**	2.656**	
	(1.061)	(1.027)	(1.016)	(1.020)	
Prop & Customer Data					1.902*
					(1.098)
Only Prop Data (No Cust.)					4.816***
					(1.671)
log(Funding before 2019)		-0.246***	-0.168*	-0.194*	-0.173*
		(0.093)	(0.100)	(0.099)	(0.096)
log(Age)			28.424***	28.524***	37.688***
			(9.907)	(10.067)	(10.743)
log(Age^2)			-13.337***	-13.333***	-17.331***
			(4.680)	(4.757)	(4.962)
SF				3.496	3.467
				(2.163)	(2.144)
Observations	150	150	150	150	150
Adj. R2	0.0291	0.0823	0.0955	0.111	0.127

Notes: * p<0.1, ** p<0.05, *** p<0.01. Coefficients are estimated using OLS regression, and include robust standard errors, in parentheses below the coefficient. Models includes CEM, based on firm age (cont.), employment size (buckets) and regions (buckets), dropping 9 firms. All these models do not use Heckman's selection procedure, controlling with IMR.

Figure A8 - Heterogeneous Effects: Advantage from Data



Note: Confidence intervals at the p<0.1 level.

Table A9 - Proprietary Data & Funding Growth Dummy (CEM)

	(1)	(2)	(3)	(4)	(5)	(6)
			Funding Growth Increase (2019, dummy)			
		OLS			Probit	
Any Prop Data	0.226***	0.226***	0.251***	0.689***	0.683***	0.949***
	(0.058)	(0.058)	(0.058)	(0.242)	(0.245)	(0.260)
IMR		0.153	0.371		-0.208	1.298
		(0.306)	(0.330)		(1.056)	(1.177)
log(Funding before 2019)			-0.004			-0.022
			(0.006)			(0.020)
log(Age)			1.384*			25.769
			(0.790)			(18.644)
log(Age^2)			-0.593			-11.047
			(0.367)			(7.993)
SF			0.312**			1.523***
			(0.138)			(0.441)
CEM Weighted						
Observations	150	150	150	159	159	159
Adj. R2	0.0731	0.0687	0.0922			

Notes: * p<0.1, ** p<0.05, *** p<0.01. CEM drops 9 firms in the OLS models. Since DV is binary, we estimate coefficients using OLS and Logit specifications. Robust standard errors are in parentheses below the coefficient. All models include Heckman's selection procedure, controlling with IMR.