On resource complementarity among startups, accelerators and financial investors: a large-scale analysis of sorting and value creation

Online Appendixes

Contents

List of Tables

List of Figures

Appendix 1 Formal model

The value of a business idea depends not only on the potential of the idea itself but also on the startup team's resources and capabilities to execute it. A common theme among practitioners in the field of high-tech entrepreneurship is that nascent ventures face two prominent risks: business risk and technology risk (Blank, 2009). Business risk is the concern that the company has a sound business model, has the capacity to generate profits, and will find enough customers before running out of funding. Technology risk asks if the appropriate technology is in place to bring the core startup product or service to market successfully. Both of these concerns are vital for an entrepreneurial venture, and both have the potential to cause a nascent business to fail (Blank, 2009). Entrepreneurs are endowed with some technological and business knowledge they can use to reduce the aforementioned risks. Combining these elements together, we can represent the value of a new startup formally as:

$$E(V) = Vt(x)b(y) - F \tag{1}$$

The parameter V represents the underlying value of the idea and $t(x)$ and $b(y)$ the associated probabilities of realization. The first probability is an inverse measure of technological risk. The second probability is an inverse measure of business risk. Note that $0 < t(x) < 1$ and $0 < b(y) < 1$. Entrepreneurs are endowed with some initial technological and business knowledge, x and y respectively, that can reduce venture risk. Business knowledge y reduces business risk, while technological knowledge x limits technology risk. The initial endowment depends on the entrepreneur's education and skills, as well as the resources provided by partners. We can write the risk functions as: $t(x) = x^{\alpha}$ and $b(y) = y^{\beta}$. This functional form ensures that both technological and business resources are necessary for startup success.

The parameter F captures the fixed cost necessary to launch the business. We assume that entrepreneurs need the help of a seed investor to cover this cost and launch their startup. Investors can cover a fraction $a > 0$ of the sunk cost F in exchange for a fraction $(1 - z)$ of startup value. We can represent the profit function of a generic entrepreneur i raising seed funds from investor j as:

$$\pi_{ij} = zVt(x_i)b(y_i) - (1 - a_j)F \tag{2}$$

Symmetrically, we can represent the profit function of a generic investor j investing in startup i as:

$$\pi_{ji} = (1 - z)Vt(x_i)b(y_i) - a_jF \tag{3}$$

The total economic value generated by the entrepreneur-investor match is simply the sum of the two profits, which coincides with startup value as outlined in equation (1):

$$V_{ij} = \pi_{ij} + \pi_{ji} \tag{4}$$

In our framework, there are two types of seed investors: primarily financial investor P (e.g., a wealthy individual) and accelerator A. For simplicity's sake, we assume that primarily financial investors can cover a fraction $a_P > 0$ of the sunk cost F, while accelerators only cover a fraction $a_A > 0$, with $a_P > a_A$. However, accelerators provide value in the form of additional "business resources," denoted by Δy. This latter variable captures the increase in the entrepreneur's business skills after attending the accelerator program. As detailed in the previous section, this effect can be driven directly by accelerator training or indirectly by interaction with peers and mentors. Thus, entrepreneurs can effectively trade initial financial resources for additional business knowledge.

Under the assumption that both technological and business knowledge have decreasing returns (thus $\alpha, \beta < 1$), we can outline the main proposition of our theoretical framework:

Core Proposition. *Technological knowledge of an entrepreneurial team and accelerator support are complements in value creation. Conversely, business knowledge and accelerator support are substitutes in value creation.*

Our proposition has some important corollaries. The first corollary is related to assortative matching, while the second is related to the realized value when the right actors match. We discuss these corollaries in the next paragraphs.

Corollary 1. Startup - investor sorting in a competitive market

Seed investing can be characterized as a two-sided matching between entrepreneurs and investors. The investor's ability to select the most promising startups (and thus generate a large return on investment) is constrained by the startup's willingness to select that specific investor, and vice versa. In this section, we analyze the baseline scenario in which we have only two startups and two investors with heterogeneous resources looking for the perfect match. Startups can select only one investor and investors can select only one startup. Sorting is a rational process in which agents maximize their payoff function, identifying the best possible match among the available options. The conclusions of this baseline scenario can be extended to large markets without losing generality.

We start by characterizing the startups. We assume the two startups have similar underlying ideas V but different founding teams. Startup T is launched by a team of scientists

with deep technological knowledge but no business knowledge. Thus, startup T's technological knowledge endowment is $x = x_{high}$, whereas its business knowledge endowment is $y = 0$. Startup B is launched by a team with more balanced resources. Startup B's technological knowledge endowment is $x = x_{low}$, whereas its business knowledge endowment is $y = y_{low}$. These startups seek the support of the two different seed investors: primarily financial investor P and accelerator A.

The equilibrium allocation of such a matching market is the following:

Hypothesis 1. *In the competitive market equilibrium, the entrepreneurial team with specialized technology knowledge pairs up with the accelerator whereas the entrepreneurial team with already good business knowledge pairs up with the primarily financial investor.*

Proof: The allocation described in Hypothesis 1 is an equilibrium allocation if its value is more than or equal to the total value generated by any alternative pairing. This equilibrium condition is known as the local value maximization condition. In our context:

$$V_{TA} + V_{BP} \geq V_{TP} + V_{BA} \tag{5}$$

Plugging in the value functions of the different pairings, we can easily show that the above condition is always satisfied as long as $(y_{low})^{\alpha} + (\Delta y)^{\alpha} > (y_{low} + \Delta y)^{\alpha}$. This latter equation is satisfied in the presence of decreasing returns to scale of business resources y, or when $\alpha < 1$.

Allocations violating condition 5 are not stable over time because rational agents would adjust their partner's compensation to attract the value-maximizing partner. For example, assume the startup with specialized technological knowledge is accidentally paired with the primarily financial investor. Because the investor can create more value when paired with the other startup, they have an incentive to increase the startup's share of the pie (parameter z) to attract the alternative investment option. Thus, inefficient allocations should dissolve over time as the system moves towards the equilibrium.

Corollary 2. Value creation in matched vs mismatched pairings

Hypothesis 2. *The startup with specialized technology knowledge is able to generate more value in the scenario in which it is supported by the accelerator than in the counterfactual scenario in which the startup is supported by the primarily financial investor.*

Hypothesis 3. *The value-adding effect of the accelerator is lower in the counterfactual sce-*

nario in which it supports the startup with existing business knowledge than in the scenario in which it supports the startup with no business knowledge.

Proof: Hypothesis 2 derives from the fact that $V_{TA} - V_{TP} > 0$. Indeed, because both technological and business resources are necessary for startup success, $V_{TP} = 0$. Hypothesis 3 derives from the local value maximization condition we discussed before. Indeed, by rearranging inequality 5 we obtain: $V_{BA} - V_{BP} \leq V_{TA} - V_{TP}$. This is the formal definition of our Hypothesis 3.

Appendix 2 Geographical Distribution of Startups

The CB database provides detailed location information for startups, including country, state, and city levels. While this information was comprehensive for the vast majority of companies, there were 67 cases where location details were not directly available. For these instances, we employed a manual coding process, utilizing ancillary information found within the CB database, such as area codes from phone numbers, or conducting targeted internet searches to ascertain their locations.

Table A2.1: Distribution of startups across geographical areas

Geographical area	Number of startups	Percentage (%)
Africa/Middle East	212	3.1
Asia	268	3.9
Australia and New Zealand	167	2.4
Canada	200	2.9
Eastern Europe	331	4.9
United Kingdom	534	7.8
India	233	3.4
Israel	77	1.1
South America	533	7.8
United States	2933	43.0
Western (Continental) Europe	1331	19.5
Total	6819	100.0

Consequently, we successfully determined the locations for 6,819 out of the 6,824 startups in our sample. Given the diverse international composition of the startups, we categorized them into eleven broader geographical areas. These areas were defined based on economic and cultural homogeneity to facilitate a coherent analysis of geographic distribution. Table A2.1 outlines the distribution of startups across these defined geographical areas. Table A2.2 reports the distribution of US startups across states.

Table A2.3 presents a cross-tabulation of startups and investors categorized by geographical area. In this table, the rows report the locations of startups, while the columns refer to the locations of their investors. Notably, a significant portion of startups in our sample are concentrated in specific regions: 43.0% are based in the USA, and 19.5% are located in Western Europe. Furthermore, substantial overlap in geographical locations for startups and their investors is observed, with 36.5% of all startups in our sample being located in the USA and also having investors from the same region. That means that around 85% (i.e., 36.5/43.0) of startups located in the USA received funding from investors located in the

Table A2.2: Distribution of U.S. Startups by State

State	Startups	%	State	Startups	%
Alaska	1	0.0	Montana	4	0.1
Alabama	7	0.2	North Carolina	28	1.0
Arkansas	8	0.3	North Dakota	1	0.0
Arizona	16	0.5	Nebraska	19	0.6
California	1088	37.1	New Hampshire	5	0.2
Colorado	95	3.2	New Jersey	19	0.6
Connecticut	22	0.8	New Mexico	3	0.1
District of Columbia	34	1.2	Nevada	14	0.5
Delaware	12	0.4	New York	419	14.3
Florida	48	1.6	Ohio	90	3.1
Georgia	38	1.3	Oklahoma	7	0.2
Hawaii	16	0.5	Oregon	19	0.6
Iowa	13	0.4	Pennsylvania	131	4.5
Idaho	4	0.1	Rhode Island	19	0.6
Illinois	79	2.7	South Carolina	8	0.3
Indiana	18	0.6	Tennessee	60	2.0
Kansas	6	0.2	Texas	98	3.3
Kentucky	25	0.9	Utah	18	0.6
Louisiana	6	0.2	Virginia	32	1.1
Massachusetts	159	5.4	Virgin Islands	1	0.0
Maryland	42	1.4	Washington	79	2.7
Maine	4	0.1	Wisconsin	32	1.1
Michigan	24	0.8	*Not available*	6	0.2
Minnesota	14	0.5	**Total**	**2933**	**100**
Missouri	38	1.3			
Mississippi	4	0.1			

Table A2.3: Cross-tabulation of startups and investors by geographical area

	Investor Location												
	1	2	3	4	5	6	7	8	9	10	11	NA	Total
Startup Location													
1 Africa	2.2	0.0	0.0		0.0	0.2	0.0		0.0	0.3	0.2	0.1	3.1
2 Asia	0.0	2.8	0.1			0.1	0.1		0.2	0.4	0.0	0.2	3.9
3 Australia	0.0		2.1		0.0	0.0			0.0	0.1	0.0	0.1	2.4
4 Canada	0.0	0.0		2.0	0.0	0.1			0.1	0.7	0.0	0.0	2.9
5 Eastern Europe	0.0	0.1		0.0	3.4	0.2	0.0	0.0	0.1	0.3	0.5	0.2	4.9
6 United Kingdom	0.1	0.0	0.0		0.2	5.2	0.0	0.0	0.2	0.6	1.2	0.2	7.8
7 India	0.1	0.1				0.0	2.3		0.2	0.4	0.1	0.2	3.4
8 Israel		0.1				0.0		0.4		0.2	0.4		1.1
9 South America		0.2		0.0	0.0	0.0			5.6	0.4	1.3	0.3	7.8
10 USA	0.2	0.3	0.2	0.1	0.5	0.5	0.1	0.1	2.0	36.5	1.1	1.4	43.0
11 Western Europe	0.1	0.1	0.1	0.0	0.2	1.3	0.0		0.5	0.6	16.3	0.4	19.5
Total	2.7	3.8	2.6	2.2	4.4	7.6	2.6	0.5	9.0	40.3	21.2	3.2	100.0

Table A2.4: Distance (in km) between startups and investors

Distance startup-investor (km)	No. of startups	%
0	2348	35.7
0-100	1154	17.6
100-500	701	10.7
500-1000	406	6.2
1000-2000	543	8.3
2000-5000	575	8.8
5000+	844	12.8
Total	6571	100.0

same country. Similarly, for Western Europe, around 84% (i.e., 16.3/19.5) of the startups in our sample located in that region received funding from investors from the same region. This pattern underscores a pronounced trend of geographical congruence between startups and their funding sources. The high degree of geographical overlap indicated in Table A2.3 reflects the propensity for investors to engage with startups within familiar or easily accessible regions, suggesting that local networks and knowledge play crucial roles in investment decisions.

Further corroborating this trend, Table A2.4 details the distribution of startups by geographical distance from their investors, revealing that over 50% of all investments occur within 100 km. This proximity underscores the preference among investors for companies located nearby, likely driven by familiarity and accessibility. Figure A2.1 visually represents this distribution, differentiating between accelerated and non-accelerated startups. The distribution patterns between the two groups are remarkably similar, with a notable concentration of investments at relatively short distances. Interestingly, accelerated startups show a slightly higher propensity to be located at both very close and very distant locations from their investors. The distribution patterns of accelerators and primarily financial investors are remarkably similar, with a notable concentration of investments at relatively short distances. Interestingly, accelerated startups show a slightly higher propensity to be located at both very close and very distant locations from their investors.

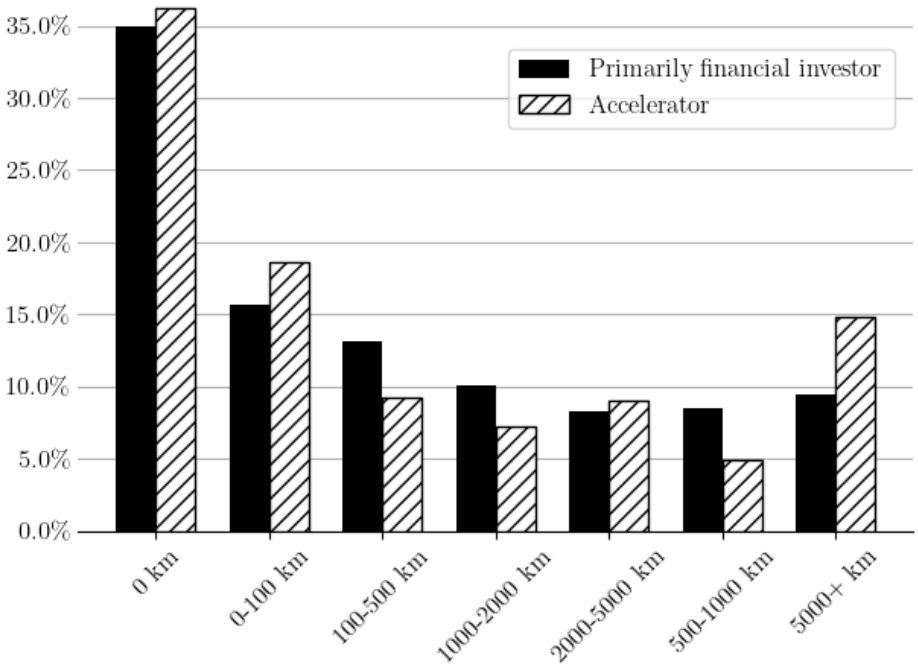


Figure A2.1: Distribution of firms by geographical distance from the investor

Appendix 3 Additional Analyses

Table A3.1: Matched sample using alternative sample selection criteria

Variable	**Accelerator**		**Primarily Financial Investor**		**Mean Diff**
	Mean	**SD**	**Mean**	**SD**	
Pr(Accelerator)	0.677	0.181	0.675	0.181	0.002
Pure tech education	0.364	0.481	0.352	0.478	0.012
Business education	0.486	0.500	0.516	0.500	-0.030
Venture age	1.363	1.956	1.288	1.596	0.075
Team age	34.508	7.312	34.384	7.177	0.124
Team size	1.967	1.025	1.958	1.027	0.009
Female incidence	0.143	0.301	0.142	0.295	0.001
Serial founder	0.561	0.497	0.540	0.499	0.021
Academic founder	0.184	0.388	0.180	0.384	0.004
MSC	0.456	0.498	0.473	0.500	-0.017
PhD	0.226	0.418	0.209	0.407	0.017
MBA	0.200	0.400	0.206	0.405	-0.006
Work experience	8.202	6.803	8.176	6.566	0.026
Top university	0.244	0.430	0.216	0.412	0.028
Top employer	0.186	0.389	0.194	0.396	-0.008
Geographical distance	0.075	1.127	-0.091	0.857	0.166
Tech similarity	0.188	0.285	0.162	0.269	0.026
Commerce	0.241	0.428	0.262	0.440	-0.021
Software	0.181	0.385	0.185	0.388	-0.004
Media & entertainment	0.096	0.295	0.088	0.283	0.008
Hardware	0.071	0.257	0.068	0.252	0.003
Mobile apps	0.070	0.255	0.065	0.247	0.005
Data Analytics	0.077	0.267	0.074	0.262	0.003
Fintech	0.079	0.270	0.073	0.260	0.006
Biotech	0.052	0.222	0.056	0.230	-0.004
Sales & marketing	0.018	0.133	0.021	0.143	-0.003
Green tech & energy	0.051	0.220	0.052	0.222	-0.001
Internet services	0.041	0.198	0.034	0.181	0.007
Design & fashion	0.024	0.153	0.025	0.156	-0.001
Africa/Middle East	0.012	0.109	0.015	0.122	-0.003
Asia	0.036	0.186	0.035	0.184	0.001
Australia and New Zealand	0.012	0.109	0.013	0.113	-0.001
Canada	0.023	0.150	0.019	0.137	0.004
Eastern Europe	0.029	0.168	0.028	0.165	0.001
India	0.027	0.162	0.031	0.173	-0.004
South America	0.015	0.122	0.012	0.109	0.003
United States	0.376	0.485	0.353	0.478	0.023
Western (Continental) Europe	0.449	0.498	0.475	0.500	-0.026
Expected joint value	0.974	0.891	0.992	0.739	-0.018
Amount first round (log)	11.067	1.369	11.939	1.188	-0.872*
Amount first round ('000s USD)	169.555	272.287	285.288	319.147	-115.733*
Observations	**1000**		**1000**		

Note: * indicates statistical significance at $p < 0.05$.

Table A3.2: Impact of Accelerators on Pure Tech and Business Ventures
Alternative Matched Sample

Pure Tech Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	-0.253	-0.215	-0.144	-0.025
	(0.016)	(0.000)	(0.000)	(0.153)
Pure tech education	-0.045	-0.036	-0.054	0.012
	(0.719)	(0.258)	(0.069)	(0.546)
Accelerator × Pure tech education	0.539	0.038	0.101	0.009
	(0.002)	(0.406)	(0.016)	(0.742)
Constant	1.508	0.673	0.358	0.113
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	2,000	2,000	2,000	2,000
R-squared	0.008	0.042	0.017	0.002

Business Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	0.143	-0.213	-0.058	-0.016
	(0.232)	(0.000)	(0.042)	(0.416)
Business education	0.225	0.026	0.096	0.023
	(0.060)	(0.402)	(0.001)	(0.248)
Accelerator × Business education	-0.398	0.024	-0.097	-0.009
	(0.019)	(0.583)	(0.015)	(0.740)
Constant	1.377	0.647	0.289	0.105
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	2,000	2,000	2,000	2,000
R-squared	0.003	0.043	0.020	0.002

Notes: The dependent variable in column (1) is the log of the incremental funding amount. p-values in parentheses.

Table A3.3: Impact of Accelerators on Pure Tech and Business Ventures
Alternative Control Group (Venture Capitalists and Business Angels)

Pure Tech Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	-0.058	-0.078	-0.079	-0.011
	(0.576)	(0.005)	(0.001)	(0.457)
Pure tech education	-0.160	-0.041	-0.043	-0.021
	(0.223)	(0.238)	(0.155)	(0.272)
Accelerator × Pure tech education	0.463	0.105	0.083	0.024
	(0.011)	(0.027)	(0.048)	(0.367)
Constant	1.452	0.542	0.301	0.095
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1,988	1,988	1,988	1,988
R-squared	0.004	0.004	0.005	0.001

Business Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	0.189	0.002	-0.033	0.003
	(0.110)	(0.938)	(0.233)	(0.875)
Business education	0.321	0.105	0.076	0.034
	(0.008)	(0.001)	(0.006)	(0.058)
Accelerator × Business education	-0.148	-0.083	-0.029	-0.009
	(0.388)	(0.065)	(0.461)	(0.709)
Constant	1.237	0.476	0.248	0.071
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1,988	1,988	1,988	1,988
R-squared	0.005	0.007	0.009	0.003

Notes: The dependent variable in column (1) is the log of the incremental funding amount. p-values in parentheses. The analysis is conducted using the original matched sample.

Table A3.4: Impact of Accelerators on Pure Tech and Business Ventures
Alternative Accelerator Definition

Pure Tech Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
AcceleratorPB	0.023	-0.079	-0.048	0.003
	(0.823)	(0.002)	(0.038)	(0.858)
Pure tech education	-0.023	-0.017	-0.014	-0.007
	(0.852)	(0.598)	(0.624)	(0.702)
AcceleratorPB × Pure tech education	0.340	0.082	0.052	0.039
	(0.053)	(0.069)	(0.191)	(0.150)
Constant	1.356	0.548	0.279	0.094
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	2,206	2,206	2,206	2,206
R-squared	0.004	0.005	0.002	0.002

Business Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
AcceleratorPB	0.275	-0.000	-0.009	0.037
	(0.014)	(0.988)	(0.710)	(0.031)
Business education	0.413	0.111	0.085	0.053
	(0.000)	(0.000)	(0.001)	(0.004)
AcceleratorPB × Business education	-0.258	-0.104	-0.037	-0.042
	(0.123)	(0.015)	(0.324)	(0.098)
Constant	1.150	0.489	0.234	0.066
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	2,206	2,206	2,206	2,206
R-squared	0.008	0.009	0.007	0.005

Notes: The dependent variable in column (1) is the log of the incremental funding amount. p-values in parentheses. The analysis is conducted using the original matched sample.

Table A3.5: Impact of Accelerators on Pure Tech and Business Ventures
Alternative Dependent Variables

Pure Tech Ventures

	(1) Total Funding Rounds	**(2)** Total Employees	**(3)** Total Revenue	**(4)** Successful Exit	**(5)** Failed
Accelerator	-0.036	-0.220	-0.237	-0.002	0.025
	(0.217)	(0.000)	(0.507)	(0.822)	(0.156)
Pure tech education	-0.048	-0.655	-1.149	-0.005	-0.030
	(0.184)	(0.000)	(0.008)	(0.704)	(0.153)
Accelerator × Pure tech education	0.125	0.373	0.935	0.012	0.008
	(0.013)	(0.000)	(0.145)	(0.490)	(0.794)
Constant	0.855	4.050	0.116	0.057	0.182
	(0.000)	(0.000)	(0.633)	(0.000)	(0.000)
Observations	3,052	1,784	337	3,052	3,052
R-squared			0.021	0.000	0.002

Business Ventures

	(1) Total Funding Rounds	**(2)** Total Employees	**(3)** Total Revenue	**(4)** Successful Exit	**(5)** Failed
Accelerator	0.062	0.320	0.369	0.008	0.023
	(0.063)	(0.000)	(0.394)	(0.492)	(0.240)
Business education	0.161	0.994	0.940	0.016	0.025
	(0.000)	(0.000)	(0.021)	(0.170)	(0.215)
Accelerator × Business education	-0.100	-0.667	-0.476	-0.012	0.010
	(0.036)	(0.000)	(0.426)	(0.485)	(0.711)
Constant	0.758	3.251	-0.781	0.048	0.160
	(0.000)	(0.000)	(0.012)	(0.000)	(0.000)
Observations	3,052	1,784	337	3,052	3,052
R-squared			0.019	0.001	0.003

Notes: Models (1) and (2) are estimated using a Poisson regression. Models (3), (4) and (5) use an OLS. The dependent variable in column (3) is the log of revenue. p-values in parentheses. The analysis is conducted using the original matched sample.

Table A3.6: Impact of Accelerators on Business & Serial Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	0.411	0.033	0.035	0.037
	(0.004)	(0.366)	(0.273)	(0.089)
Business & Serial	0.347	0.121	0.080	0.053
	(0.003)	(0.000)	(0.002)	(0.003)
Accelerator × Business & Serial	-0.326	-0.090	-0.080	-0.035
	(0.048)	(0.033)	(0.029)	(0.161)
Constant	1.132	0.438	0.211	0.053
	(0.000)	(0.000)	(0.000)	(0.001)
Observations	3,052	3,052	3,052	3,052
R-squared	0.005	0.007	0.004	0.003

Notes: The dependent variable in column (1) is the log of the incremental funding amount. p-values in parentheses. The analysis is conducted using the original matched sample.

Table A3.7: Impact of Most Influential Accelerators on Pure Tech and Business Ventures

Pure Tech Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	0.389	0.140	0.092	0.110
	(0.007)	(0.000)	(0.007)	(0.000)
Pure tech education	-0.086	-0.039	-0.033	-0.015
	(0.410)	(0.150)	(0.190)	(0.367)
Accelerator × Pure tech education	0.116	0.106	0.077	-0.031
	(0.623)	(0.078)	(0.166)	(0.413)
Constant	1.425	0.543	0.283	0.098
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1,873	1,873	1,873	1,873
R-squared	0.008	0.022	0.012	0.016

Business Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	0.434	0.248	0.137	0.078
	(0.005)	(0.000)	(0.000)	(0.002)
Business education	0.358	0.110	0.090	0.051
	(0.000)	(0.000)	(0.000)	(0.001)
Accelerator × Business education	-0.023	-0.144	-0.029	0.049
	(0.921)	(0.014)	(0.595)	(0.190)
Constant	1.222	0.476	0.228	0.068
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1,873	1,873	1,873	1,873
R-squared	0.016	0.030	0.019	0.024

Notes: The dependent variable in column (1) is the log of the incremental funding amount. p-values in parentheses. The analysis is conducted using the original matched sample.

Table A3.8: Impact of Accelerators on Pure Business Education and Mixed Tech-Business Education Ventures

Pure Business Education Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	-0.195	-0.018	-0.015	0.024
	(0.019)	(0.377)	(0.410)	(0.043)
Pure business education	-0.103	-0.002	-0.034	0.002
	(0.427)	(0.941)	(0.221)	(0.931)
Accelerator × Pure business education	-0.417	-0.100	-0.069	-0.079
	(0.026)	(0.028)	(0.085)	(0.004)
Constant	12.330	0.531	0.279	0.093
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	3,052	3,052	3,052	3,052
R-squared	0.009	0.004	0.005	0.005

Mixed Tech-Business Education Ventures

	(1) Incremental Funding	**(2)** Top 50% Total Funding	**(3)** Top 25% Total Funding	**(4)** Top 10% Total Funding
Accelerator	-0.267	-0.038	-0.016	0.008
	(0.003)	(0.081)	(0.398)	(0.534)
Mixed tech-business education	0.519	0.094	0.113	0.054
	(0.000)	(0.000)	(0.000)	(0.001)
Accelerator × Mixed tech-business education	0.031	0.014	-0.023	0.009
	(0.842)	(0.715)	(0.488)	(0.690)
Constant	12.132	0.498	0.234	0.075
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	3,052	3,052	3,052	3,052
R-squared	0.019	0.010	0.013	0.009

Notes: The dependent variable in column (1) is the log of the incremental funding amount. p-values in parentheses. The analysis is conducted using the original matched sample.

Appendix 4 Primary Data Analysis

To corroborate our findings and provide additional evidence of the theoretical mechanism, we performed a second analysis using data gathered through an anonymous survey. The entire survey is available at the end of this Appendix. Our final sample contained 236 unique responses[1], 56% of them from startups that are accelerated and 44% from startups that raised seed funding from primarily financial investors (control group). The breakdown of control startups by type of seed investor shows that 30% were backed by VCs, 28% by business angels, 27% by individuals, and 15% by other primarily financial investors. Based on these measures, there is no evidence of a response bias related to investor type and the subsample appears to be similar to the larger one used in the main analysis. At the end of the survey, we asked entrepreneurs in the control group to explain why they did not consider enrolling in an accelerator program. Reassuringly, the vast majority of responses (43%) reported an "exogenous reason": they were not aware of an accelerator or accelerators were not available in their city. Interestingly, 30% of the cases reported a collaboration with another seed investor as the major reason. 17% suggested they did not need any help from acceleration programs. Only, 10% were rejected applicants. This evidence is consistent with the idea that entrepreneurs frequently compare alternative seed investors and that important trade-offs exist.

Following the procedure outlined in the main analysis, we identified Pure tech teams and Business teams (i.e., ventures with at least one team member with a business background). The share of Pure tech and Business teams in the survey are in line with the descriptive statistics in the pre-matched sample, suggesting that a response bias based on educational background is very unlikely. In addition to the main independent variables, we control for entrepreneur's gender (Female dummy variable), Age bracket (<18, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, >85), immigration status (Immigrant) and Education level (1. no college education, 2. college education, 3. Bachelor or equivalent, 4. Master or equivalent, 5. PhD). The most important question in the survey is the rating of seed investor impact on startup performance using a seven-point Likert scale. The variable *Overall investor impact* reports entrepreneurs' answers to this question. As a follow-up, we asked entrepreneurs to rate how the seed investor helped their startup. This question was divided into different items encompassing all the different aspects of startup launch—from *Fundraising* to *Training on business issues*. As in the previous question, entrepreneurs were asked to rate impact using a seven-point Likert scale. All the different survey items as well as summary statistics

[1]Because the structure of our questionnaire does not allow respondents to differentiate their answers based on ventures, we excluded entrepreneurs managing more startups.

for all variables are reported in Table A4.1.

Table A4.1: Impact of Accelerators, Descriptive Statistics (Survey Data)

	Entire Sample		**Accelerator**		**Primarily Financial Investor**	
	Accelerator	Primarily Financial Investor	Pure Tech	Non-Pure Tech	Pure Tech	Non-Pure Tech
Overall investor impact	5.15	4.30	5.43	5.01	4.30	4.29
	(1.69)	(2.08)	(1.62)	(1.76)	(1.97)	(2.14)
Fundraising	4.61	6.08	4.88	4.46	6.23	6.02
	(2.25)	(1.69)	(2.22)	(2.30)	(1.63)	(1.72)
Business model validation	4.07	3.16	4.18	4.01	3.00	3.23
	(1.89)	(1.93)	(1.85)	(1.91)	(2.21)	(1.82)
Feedback on the idea	4.15	3.13	4.38	4.03	3.19	3.10
	(1.83)	(1.92)	(1.83)	(1.83)	(2.11)	(1.86)
Customer development	3.92	2.81	4.00	3.89	2.73	2.93
	(1.90)	(1.85)	(1.86)	(1.95)	(1.90)	(1.85)
Advice on operations	3.72	3.32	3.90	3.62	2.96	3.43
	(1.84)	(1.89)	(1.71)	(1.90)	(1.92)	(1.87)
Training on business issues	3.44	2.73	3.97	3.16	2.60	2.78
	(1.86)	(1.73)	(1.60)	(1.93)	(1.80)	(1.71)
Training on technical issues	2.26	1.94	2.40	2.19	1.92	1.95
	(1.55)	(1.29)	(1.52)	(1.55)	(1.28)	(1.31)
Pitching the idea	5.21	3.07	5.59	5.01	3.00	3.10
	(1.65)	(1.83)	(1.41)	(1.77)	(1.95)	(1.79)
Technology development	2.70	2.32	2.63	2.73	1.96	2.46
	(1.54)	(1.94)	(1.27)	(1.67)	(1.31)	(1.71)
Team building	2.92	2.74	3.04	2.85	2.30	2.92
	(1.69)	(1.62)	(1.66)	(1.68)	(1.46)	(1.71)
Networking with mentors	5.19	3.00	5.54	5.00	3.23	2.90
	(1.65)	(1.93)	(1.31)	(1.80)	(2.02)	(1.90)
Access to physical space	4.60	2.25	4.75	4.60	2.07	2.32
	(2.13)	(1.78)	(1.23)	(1.11)	(2.52)	(1.89)
Pure tech	0.33	0.25	1	0	1	0
	(0.48)	(0.46)	(0)	(0)	(0)	(0)
Business	0.37	0.39	0	0.55	0	0.52
	(0.48)	(0.49)	(0)	(0.49)	(0)	(0.49)
Female	0.13	0.06	0.11	0.13	0.08	0.05
	(0.33)	(0.23)	(0.30)	(0.35)	(0.24)	(0.23)
Age bracket	3.3	3.4	3.02	3.45	3.30	3.43
	(0.87)	(0.71)	(0.85)	(0.91)	(0.74)	(0.69)
Immigrant	0.25	0.14	0.29	0.23	0.23	0.10
	(0.43)	(0.34)	(0.46)	(0.42)	(0.42)	(0.30)
Education level	3.7	3.7	3.88	3.58	4.07	3.67
	(0.94)	(0.90)	(0.84)	(0.98)	(1.03)	(0.80)
Observations	133	103	44	89	26	77

Notes: The table reports the mean and the standard deviation (in parenthesis) of all variables. Variable scores are reported using a Likert scale from 1 to 7. Non-Pure Tech is the residual category composed of teams that are not pure tech, i.e., it includes business teams and all other teams.

The survey responses are consistent with the theory outlined in this study. Entrepreneurs recognize that accelerators provide more value-adding activities than primarily financial seed investors like VCs or business angels (*Overall investor impact*). Specifically, accelerators offer

more valuable feedback on the business model (*Business model validation*), idea (*Feedback on the idea*), customer development (*Customer development*), and pitching (*Pitching the idea*). Overall, as hypothesized, they provide superior training on business-related issues (*Training on business issues*) and mentoring (*Networking with mentors*). Conversely, VCs and business angels contribute with more financial resources (*Fundraising*). The picture becomes more interesting when we break down the impact of accelerator value-adding activities based on team educational background. As expected, Pure tech entrepreneurs report a stronger positive impact of the acceleration program on their startup (*Overall investor impact*). The main theoretical arguments of the paper find additional support if we look at how accelerators add value for these entrepreneurs. As shown in Table A4.1, Pure tech entrepreneurs report similar scores to other entrepreneurs in most questions related to value-adding activities, except three—*Training on business issues* ($p = .02$), *Pitching the idea* ($p = .06$), and *Networking with mentors* ($p = .08$). Conversely, we do not observe such differences between Pure tech and non-Pure tech entrepreneurs in the control group.

We test the main theoretical mechanism outlined in the paper reporting regression results using *Training on business issues* as a dependent variable. Results reported in Table A4.2 show accelerators provide more comprehensive business training than primarily financial investors (from 0.4 to 1 additional Likert scale points). This effect is stronger for Pure tech entrepreneurs (about 1 additional Likert scale point) and weaker for entrepreneurs with a Business background (from 0.6 to 0.8 less Likert scale points). Overall, these results show a strong substitution effect of acceleration training on Business teams.

Table A4.2: Accelerator Training as a Substitute of Business Knowledge (Survey Data)

Variable	**(1)** Training on Business Issues	**(2)** Training on Business Issues	**(3)** Training on Business Issues	**(4)** Training on Business Issues
Accelerator	0.387	0.525	0.967	1.169
	(0.196)	(0.090)	(0.004)	(0.001)
Pure tech	-0.181	-0.239		
	(0.669)	(0.588)		
Business			0.069	0.153
			(0.858)	(0.700)
Accelerator × Pure tech	0.990	0.858		
	(0.068)	(0.120)		
Accelerator × Business			-0.634	-0.859
			(0.213)	(0.095)
Female		0.063		-0.020
		(0.879)		(0.961)
Age bracket		-0.209		-0.238
		(0.207)		(0.140)
Immigrant		0.228		0.252
		(0.456)		(0.410)
Education level		0.709		0.863
		(0.489)		(0.397)
Constant	2.781	1.875	2.700	1.465
	(0.000)	(0.070)	(0.000)	(0.160)
Observations	216	214	216	214
R-squared	0.064	0.109	0.050	0.114

Notes: OLS regressions. p-values in parentheses.

ONLINE SURVEY

Our University is conducting a survey to better understand the impact of accelerators on startups. Data will build on entrepreneurship research and will help understanding the way in which investors support innovative businesses, with a focus on how innovators may benefit differently from different types of investors based on their background. The survey will take approximately 5 minutes to complete. Please answer as many questions as possible.

Note: By responding to this survey, you personally consent to have your responses used in the research study. These responses represent your personal views and opinions. You also understand that this survey will not be asking you to reveal any confidential business information. Your answers will be used only by the researchers at our university, will be aggregated, and anonymized in any publication.

Q1. Have you ever taken part in an acceleration program? (Yes/No)

If Yes to Q1 (TREATED):

Q2. How long has the acceleration program lasted? (From 1 = less than a month, to 6 = more than a year)

Q3. To what extent did the acceleration program help you to launch your startup? (From 1 = not at all; to 7 = a lot)

Q4. Why did you apply to an acceleration program? Please rate the following (From 1 = not at all important; to 7 = very important):

- Raise funding
- Receive feedbacks on your business idea
- Receive help on developing a business model
- Networking

Q5. How did the acceleration program benefit your start-up? Please rate the following (From 1 = not at all important; to 7 = very important):

- Receive financing from them
- Receive help on business model validation
- Test your idea and receive feedbacks

- Receive advice on customer development
- Receive advice on business operations
- Receive advice on venture financing
- Trainings on business issues
- Trainings on technical issues
- Receive advice on pitching to investors
- Receive advice on technology and innovation
- Receive advice on recruiting people
- Receive advice on suppliers management
- Networking opportunities with investors
- Networking opportunities with other accelerated startups
- Networking opportunities with mentors
- Physical space and resources
- Other

If No to Q1 (CONTROL):

Q6. Why did you never take part in an acceleration program?

- I never thought about it
- I didn't know about the existence
- I got rejected
- I already had a good investor
- I already had a well running business
- Others: _____

Q7. Who was your first investor?

- Individual
- Business Angel
- Government Office
- Venture Capital Fund

- Corporate Venture Capital
- Start-up Competition
- Family Investment Office
- Co-working Space
- Crowdfunding Platform
- University Program
- Non-equity Program
- Venture Debt
- Private Equity Fund
- Other

Q8. To what extent did your first investor help you to launch your startup? (From 1= not at all; to 7= a lot)

Q9. Why did you choose that specific investor? Please rate the following: (From 1= not at all important; to 7= very important)

- Raise funding
- Receive feedbacks on your business idea
- Receive help on developing a business model
- Networking

Q10. How did your first investor benefit your start-up? Please rate the following (From 1= not at all important; to 7= very important)

- Receive financing from them
- Receive help on business model validation
- Test your idea and receive feedbacks
- Receive advice on customer development
- Receive advice on business operations
- Receive advice on venture financing
- Trainings on business issues
- Trainings on technical issues

- Receive advice on pitching to investors
- Receive advice on technology and innovation
- Receive advice on recruiting people
- Receive advice on suppliers management
- Networking opportunities with investors
- Networking opportunities with other accelerated startups
- Networking opportunities with mentors
- Physical space and resources
- Other

Respondent Characteristics:

Q11. Highest level of education:

- No college education
- Some college education
- BA or equivalent
- MA or equivalent
- Doctorate or equivalent
- Higher

If Q11 >1:

Q12. Field of highest degree:

- Science, Technology, Engineering and Mathematics related fields
- Business, Economics or Law related fields
- Other: _____

Q13. Field of highest degree of the others founders

- All Science, Technology, Engineering and Mathematics related fields
- All Business, Economics or Law related fields
- A mix of the previous two
- Other _____

- They didn't own a degree
- There were no other founders

Q14. What was your age at the time of venture founding?

- Under 18
- 18 – 24
- 25 – 34
- 35 – 44
- 45 – 54
- 55 – 64
- 65 – 74
- 75 – 84
- 85 or older

Q15. What is your gender?

- Male
- Female
- Other

Q16. Where you originally from a different country compared to the one where you founded the venture? (Yes/No)

Q17. Have you previously founded a start-up? (Yes/No)

Appendix 5 Sample Selection

As noted in the manuscript, we extracted information from the Crunchbase (hereafter referred to as CB) database through the dedicated RESTful API. We focused on startups that received their first funding round between 2004 and 2018 (inclusive) from the universe of all firms contained in CB. We further narrowed our selection to startups that received an amount less than or equal to US$150,000 in their first funding round. We excluded ventures for which the type of investors in the first funding round was not disclosed, as well as those funded by companies or pension funds. These criteria led us to identify an initial sample of 12,759 firms.

Figure A5.1 offers a comprehensive overview of the principal steps involved in constructing our final sample. The subsequent sections will delve into a detailed commentary on these steps, elucidating the rationale behind each decision and the methodologies employed to refine our sample.

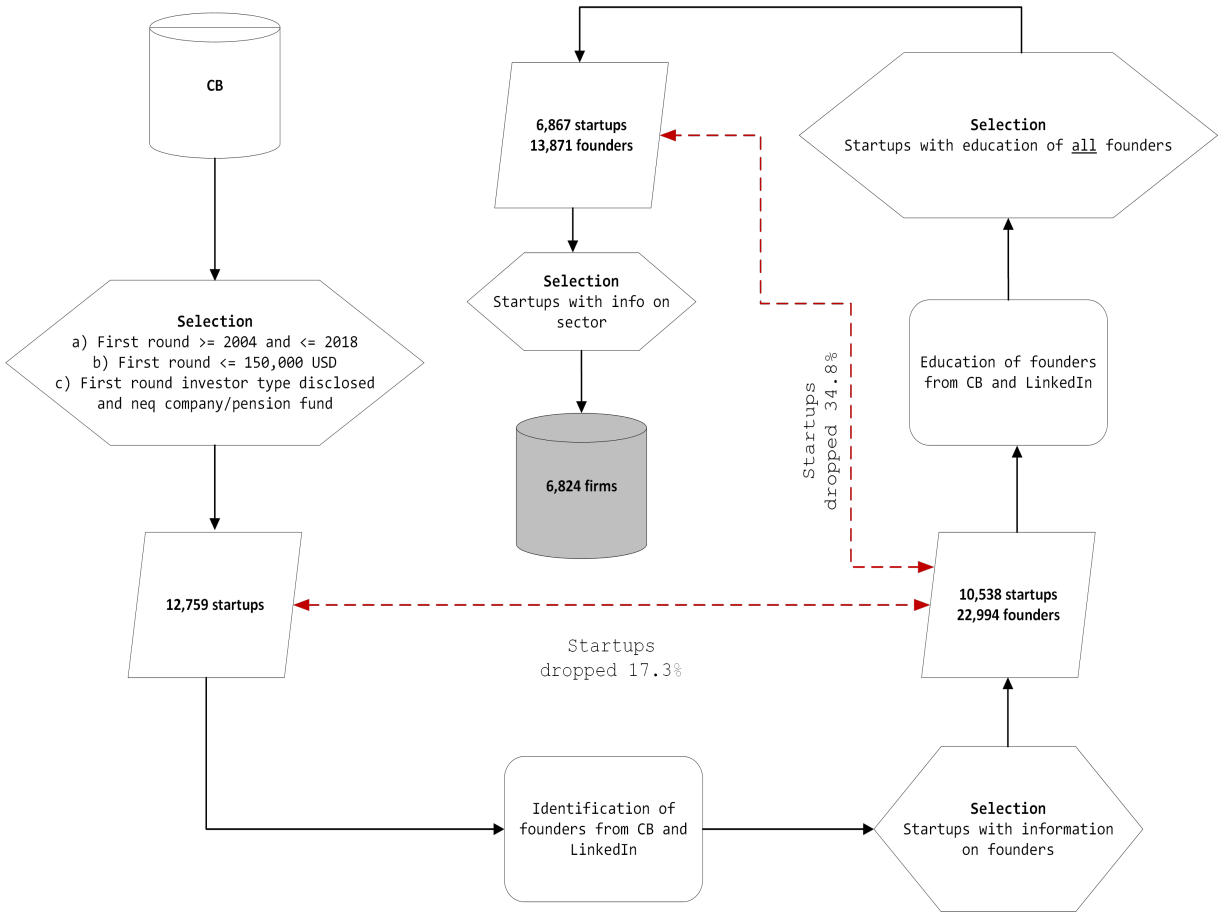


Figure A5.1: Sample construction: flowchart

5.1 Identification of Startup Founders

To identify the founding members of the selected startups, we utilized information from the *jobs* and *people* tables, retrieved via the RESTful API from the CB database. Specifically, we identified individuals affiliated with any of the selected startups (either currently or in the past) who reported their job titles as *founder* or *co-founder*. We combined this information with data from the CB summary pages of all the ventures in our sample (see Figure A5.2 for an example).

Combining these two pieces of information is crucial for the accurate identification of all founders. In the example shown in Figure A5.2, one of the two founders, Dmitry Kaigorodov, is listed with the job title of CEO in the *jobs* and *people* tables, yet he is recognized as a co-founder on the summary page.

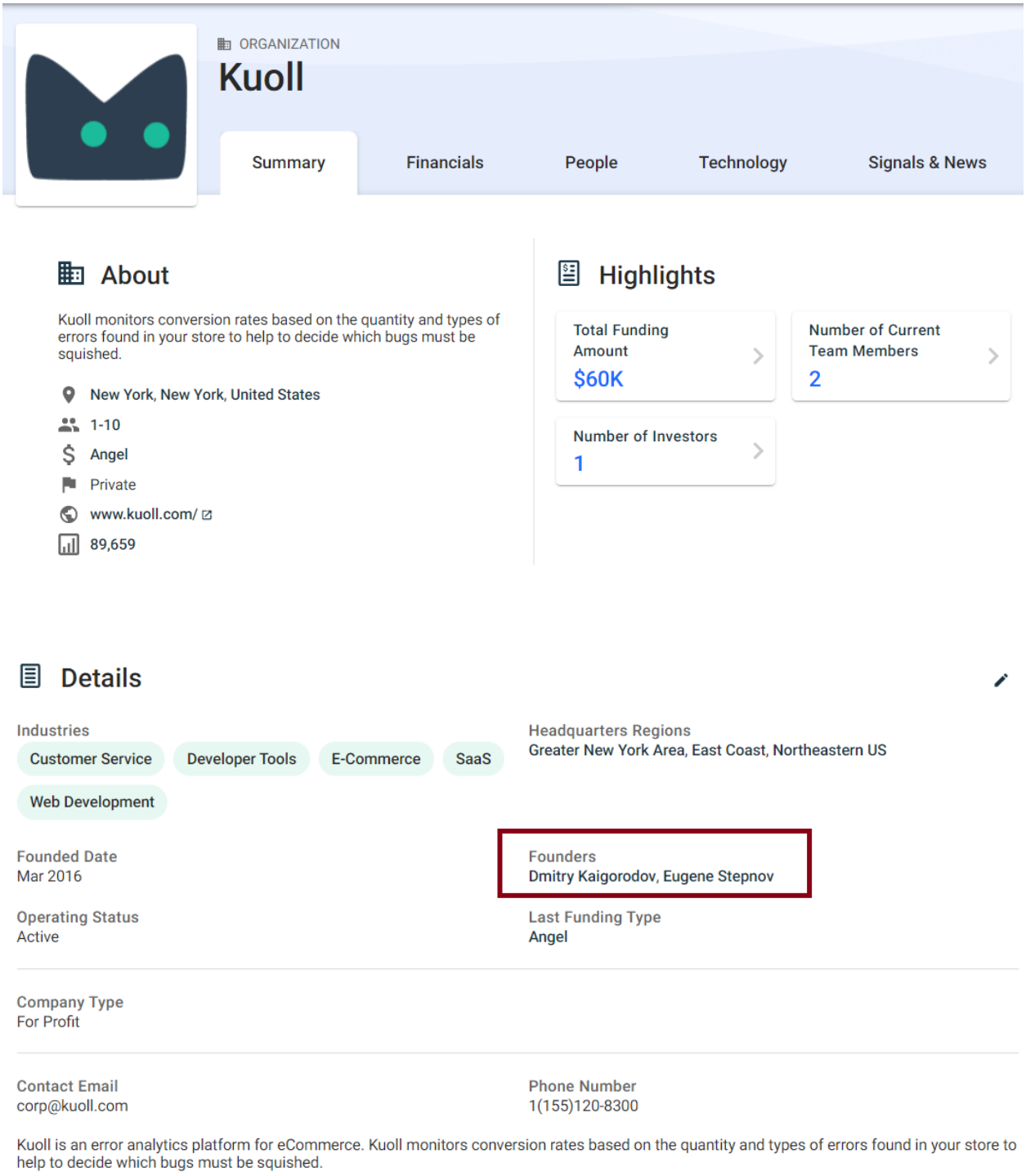
The process of identifying startup founders through the data available in CB introduces a second challenge—not all *true* founders of a startup necessarily have a CB profile. As a result, the names, affiliations, and job titles of some true founders may be absent from the CB database. To mitigate this issue, we utilized information from LinkedIn. Specifically, for each startup in our sample, we collected the public profiles of all employees with a reported affiliation to that startup. Within this group, we identified as founders those individuals who, on their LinkedIn profiles, listed *founder* or *co-founder* as their job title. We employed two approaches for this purpose—either by reviewing the social and professional networks often linked on CB profiles, specifically Facebook, Twitter, and LinkedIn, or, for founders without a social media presence on CB, by performing a manual Google search.

Out of the 12,759 startups in the initial sample, our methodology helped in identifying the founders of 10,538 ventures. In total, we identified 22,994 *distinct* founders[2]. The distribution of these founders, based on the source of the information, is detailed in Table A5.1.

These results align closely with a recent benchmarking exercise conducted by Retterath and Braun (2020), who compared information from eight commonly used VC databases, including CB, across 339 actual VC financing rounds involving 396 investors and 108 different companies, predominantly in Europe. They found that CB, along with Pitchbook, offers the best coverage regarding the number and educational backgrounds of founders. Their analysis indicates that CB reports 63% of all *true* founders, as verified by funding contracts and original documentation, which is less than the 76% reported in our study. However, it is important to note the differences in the scope of the two studies: Retterath and Braun (2020) primarily focused on European companies and a highly selected sample of firms that received

[2]It is important to note that some of these individuals may have founded more than one startup.

Figure A5.2: Summary page of a startup on the CB website



`https://www.crunchbase.com/organization/kuoll` (consulted on May 1st 2021).

Table A5.1: Origin of information on founders

Origin of Information	**Number of Founders**	**%**
(a) Individuals affiliated with a startup and reporting founder or co-founder as job title in the CB database	17,526	76.2
(b) Individuals affiliated with a startup who do not report founder or co-founder as job title in the CB database but are listed as part of the founding team on the CB website	2,968	12.9
(c) Individuals without a CB profile, but listed as founders (or co-founders) of the startup in their LinkedIn profiles	2,500	10.9
Total	**22,994**	**100.0**

funding exclusively from VCs, while our study encompasses a broader range of investors and includes startups that have received only one or a few funding rounds. Therefore, comparisons should be approached with caution, taking these contextual differences into account.

Our methodology for identifying founders using CB and LinkedIn closely mirrors that of Roche, Conti, and Rothaermel (2020), who successfully identified founding teams for 1,790 out of 2,064 companies in their sample, achieving an identification rate of 87%. In contrast, our study identified founding teams for 10,538 startups out of an initial pool of 12,759, yielding an 82.6% success rate. Given the broader geographical and sectoral coverage of our research compared to that of Roche, Conti, and Rothaermel (2020), these figures further validate the robustness and reliability of our sample construction process.

It is conceivable that some *true* founders may not have been detected in our study. Without direct access to company files and the individuals involved in the founding process, obtaining a more accurate representation of the founding teams would be challenging. Nevertheless, if any misrepresentations were to occur randomly across different types of investors, including accelerators, the overall impact on our findings would likely be minimal.

5.2 Educational background of founders

The CB API facilitates the extraction of tables containing information on the educational backgrounds of individuals registered on the platform. However, it is crucial to note that

not all profiles on CB provide details about educational attainment. For some individuals, CB offers only basic information such as name, gender, company affiliations, and job titles, without any specifics regarding their education.

Out of the 22,994 founders identified, only 6,940 had reported their educational background in their CB profile, accounting for roughly 30% of all founders. For the remaining 16,054 founders, we turned to LinkedIn. We searched for the corresponding (public) LinkedIn profiles for all 22,994 founders, successfully finding 17,947 of them, which represents slightly more than 78% of the total. This data collection process enabled us to classify the sample of 22,994 founders into four different categories based on the source of information regarding their educational background (see Table A5.2). Notably, for about 30% of the founders in our sample, the only available information on their educational background was what they had reported on their LinkedIn profile. Meanwhile, educational background information was accessible from both LinkedIn and CB profiles for 23% of the founders.

Table A5.2: Origin of information for founders' education

Origin of information	**Founders**	**%**
Founders with info from CB and LinkedIn	5,180	22.5%
Founders with information only from CB	1,760	7.7%
Founders with information only from LinkedIn	6,931	30.1%
Founders without information from any source	9,123	39.7%
Total	22,994	100.0%

Most importantly, information on the educational background was unavailable for approximately 40% of the founders in our sample. This lack of information occurred either because individuals did not report their educational background in their CB profile or because we were unable to find supplementary information from LinkedIn. As mentioned previously, not all founders possess a corresponding (public) LinkedIn profile. Moreover, some who do have a profile report incomplete information, lacking details on their educational attainment.

Given that our empirical analysis hinges on accurately identifying the educational background of *all* founders within a company, we were compelled to exclude from our sample those companies for which we could not obtain this information. This reduction narrowed our sample to 6,867 firms.

For the 13,871 founders associated with the remaining 6,867 startups, we categorized educational attainment according to: a) the type and level of degree, b) the degree's subject area, and c) the start and end years of the degree program. In the subsequent two subsections, we outline the methodology employed to classify the type and level of degree, as well as the subject area of the degree.

5.3 Checking for selection biases

As detailed above, we were forced to drop from our sample the companies for which we could not retrieve information on founders' educational background. In Table A5.3, we compare the group of companies for which we have no information on the educational background with our sample to check for potential selection biases. Overall, the two samples look remarkably similar. The first group comprises ventures with a slightly larger number of founders and a slightly lower incidence of female founders. The remaining differences should not affect our findings in a significant way.

Table A5.3: Checking selection biases

	Educational background of **all** founders	
	No	**Yes**
Amount first round	62422.8	60860.4
Total amount raised	4.356	4.558
No. of funding rounds	2.21	2.27
Company age at the first round	4.35	4.32
Year of foundation	2013.2	2013.3
Number of founders	2.52	2.03
Share of female founders	0.113	0.149
Three most represented sectors	Commerce (25.4) Software (18.6) Media (11.8)	Commerce (29.2) Software (22.4) Media (11.5)
Three most represented countries	USA (37.7) GBR (8.3) IND (3.1)	USA (43.0) GBR (7.8) ESP (3.5)
Number of startups	3,714	6,824
Startups with founders' identity	10,538	

Note: The first column refers to startups for which we had information on the names of founders, but we did not have information on the educational background of *all* founders. The second column refers to startups for which we had information on the names of founders and on the educational background of *all* founders.

Appendix 6 Coding education

6.1 Education level

To categorize the type and level of degree, we utilized the UNESCO International Standard Classification of Education (ISCED) 2011[3] as our reference classification. The ISCED 2011 framework differentiates educational attainment into nine levels. For the scope of our study, the pertinent levels were (6) Bachelor or equivalent, (7) Master or equivalent, and (8) Doctoral or equivalent. To systematically code the educational level, we employed a dictionary-based approach. Specifically, for each relevant educational level (e.g., Master), we compiled a dictionary of various (regular) expressions that might denote that level of education, such as:

{'Master': {'MSC', 'MPHIL', 'MSEE', 'MTECH', 'MMATH', etc.}}

For instances where the automated classification system, based on our specially constructed dictionary, was unable to categorize education levels, we conducted a manual review and classification. Table A6.1 displays the distribution of founders according to the highest academic degree they obtained.

Table A6.1: Highest title attained by founders

Highest Title Attained	Number of Founders	Percentage (%)
Bachelor	5,080	36.6
Master (including MBA)	5,295	38.2
PhD	1,331	9.6
Others	2,165	15.6
Total	13,871	100.0

Note: The category "Others" includes a miscellanea of titles and certifications that escape any standard classification. It also includes primary and secondary education.

The data presented in Table A6.1 show broad comparability with the figures from the benchmarking exercise conducted by Retterath and Braun (2020). In our sample, founders with a Ph.D. constitute 9.6% of all founders, slightly below the 10.9% reported in the cited study. Founders with a Master's degree represent 38.2% in our findings, compared to 46.8% in the source cited, while those with a Bachelor's degree account for 36.6%, significantly higher than the 15.2% reported by Retterath and Braun (2020). Interestingly, Retterath

[3] `https://tinyurl.com/hsz47cvn`

and Braun (2020) identified 26.5% of founders without any information on their educational background, in contrast to 15.6% in our sample. These discrepancies, particularly the higher percentage of founders with a Bachelor's degree in our study, can likely be attributed to the enhanced precision of our data, especially regarding the utilization of LinkedIn profiles.

6.2 Education field

Table A6.2: ISCED-F 2013, Fields of education and subject contents

ISCED code	ISCED area	Subject contents
01	Education	Didactics Teacher Training Education technology ...
02	Humanities	Classical Languages History Philosophy ...
03	Social Sciences	Sociology Psychology Journalism ...
04	Business	Management Science Business Finance Accounting ...
05	Natural Sciences	Biology Genomics Mathematical Biology ...
06	ICT	Computer Science Informatics Network administration ...
07	Engineering	Ceramics Electronics Materials Food processing ...
08	Agriculture	Fisheries Farming Forestry ...
09	Health	Psychiatry Physiology Pharmacy ...
10	Services	Catering Cosmetology Transport services ...

For categorizing the subject areas of degrees, we combined machine learning and natural language processing (NLP) techniques, grounding our approach in the detailed descriptions

of fields of education as outlined in the ISCED-F 2013 classification, published by UNESCO[4]. This taxonomy organizes education and training programs into ten broad areas based on the subject content of the education, in addition to a residual category (see Table A6.2). The ISCED-F 2013 classification delineates each of the ten broad fields with a comprehensive list of programs and qualifications, detailing the subject content classified under each field (see last column of Table A6.2).

The classification challenge we encountered arises from the fact that the degree titles reported in the CB and LinkedIn profiles do not always directly match the subject contents listed in the ISCED-F 2013 dictionary, as depicted in Table A6.2. This discrepancy complicates the categorization process. For instance, consider the degree title found on the CB profile of a founder in our sample and reported in Table A6.3. The issue is that *bioinformatics* is not reported as subject content area in any of the education fields classified by ISCED and reported in Table A6.2.

Facing the challenge that certain degree titles, such as *bioinformatics*, do not align with any subject content areas defined by the ISCED-F 2013 classification, our objective was to develop an algorithm capable of accurately predicting the most suitable field of education for these degrees, aligning them with one of the ten broad fields outlined in the ISCED-F 2013 framework.

To achieve this, we used a combination of machine learning and NPL techniques. We describe each of them in turn.

6.3 Machine learning

We employed three supervised machine-learning algorithms designed for short text categorization, utilizing the detailed list of programs and qualifications from the ISCED-F 2013 classification as our *training set*. These algorithms are based on *word-embedding cosine similarity classifiers*, specifically tailored for processing and analyzing textual data[5].

The first algorithm utilizes the Word2Vec model, which is among the most prevalent word-embedding techniques. Developed by Mikolov, Sutskever, et al. (2013) and Mikolov, Chen, et al. (2013), Word2Vec transforms words into vectors that encapsulate semantic meaning within an n-dimensional vector space. This method allows for the grouping of semantically similar words based on their co-occurrence within large text corpora. We applied Google's pre-trained Word2Vec model, which contains vectors for 3 million words and phrases derived

[4] `https://tinyurl.com/mrx9hdcw`

[5] For this analysis, we utilized the `shorttext` Python library, available at `https://shorttext.readthedocs.io/en/latest/`.

Table A6.3: Example of education profile

Variable Name	**Description**	Example
uuid	Unique identifier of degree	829cd8ca-80ac-253a-c85b-587832b83df7
name	Degree name	Ph.D. Bioinformatics @ University of California, Santa Cruz
person_uuid	Unique identifier of person	caf8dbdb-5380-a162-3124-ec93d9d79754
person_name	Person name	Charles Vaske
institution_uuid	Unique identifier of educational institution	f2d37262-3642-64b7-e584-08a04c5698b4
institution_name	Educational institution name	University of California, Santa Cruz
degree_type	Type of degree	Ph.D.
subject	Educational subject	Bioinformatics
started_on	Starting date of education	2003
completed_on	Completion date of education	2009

from a dataset of approximately 100 billion words from Google News[6], with each word represented as a 300-dimensional vector.

For instance, using the Google News dataset's pre-trained Word2Vec model, the term *bioinformatics* yields the top five semantically similar words: [('genomics', 0.72), ('proteomics', 0.71), ('computational_biology', 0.71), ('informatics', 0.71), ('computational_chemistry', 0.69)], showcasing the model's ability to capture and reflect the semantic proximity of related terms[7].

The second algorithm we applied is the GloVe model (Global Vectors for Word Representation), introduced by Pennington, Socher, and Manning (2014). While similar to Word2Vec in converting words into vectors, GloVe distinguishes itself by incorporating both local and global statistical information from the corpus to generate word vectors, thus enhancing the model's ability to capture broader contextual meanings. We utilized two GloVe models pre-trained on distinct corpora: one on Wikipedia and Gigaword with a vocabulary of 400,000 words, and another on the Common Crawl dataset, encompassing 1.9 million words and phrases. Both models feature 300-dimensional vectors for each word, facilitating a nuanced analysis of textual data akin to that achieved with Word2Vec.

Leveraging these pre-trained models, we calculated the cosine similarity between the degree title in question (e.g., *bioinformatics*) and the listed subject contents across each of the eleven fields of education as delineated in Table A6.2. The degree title was allocated to the field exhibiting the highest cosine similarity. For the specific case of *bioinformatics*, all models classified this degree title under the natural sciences field (see Table A6.4).

6.4 Soft Jaccard Score

Beyond the word-embedding models previously described, our analysis also incorporated the Soft Jaccard Score, a sophisticated metric assessing the edit distance between two sets of tokens, as proposed by Russ et al. (2016). The Soft Jaccard Score extends the traditional Jaccard similarity measure to account for partial matches between elements of two sets. This is particularly useful for comparing lists of textual tokens, where variations in spelling or form can still indicate a meaningful similarity.

Given two sets of tokens, T_1 and T_2, the calculation of the soft Jaccard score involves the following steps:

1. Calculate the similarity between any two tokens t_1 and t_2 as the maximum of two

[6]Google's pre-trained Word2Vec model is available for download at `https://code.google.com/archive/p/word2vec/`.

[7]This example demonstrates how word vectors, even for complex scientific terms, reveal closely associated fields, indicating the model's effectiveness in mapping domain-specific vocabulary.

Table A6.4: Bioinformatics: Cosine similarity scores across models

ISCED area	Word2Vec Model (Google News)	Glove Model (Wikipedia)	Glove Model (Common Crawl)
Agriculture	0.323	0.174	0.280
Humanities	0.325	0.110	0.242
Business	0.300	0.050	0.261
Education	0.265	0.084	0.273
Engineering	0.405	0.167	0.297
Health	0.417	0.267	0.380
ICT	0.492	0.292	0.398
Natural sciences	**0.712**	**0.641**	**0.735**
Services	0.260	0.008	0.188
Social sciences	0.407	0.233	0.367

measures: the Damerau-Levenshtein (DL) distance and the longest common prefix (LCP).

The Damerau-Levenshtein (DL) distance counts the edits needed (such as substitutions, insertions, deletions, or transpositions) to transform one token into the other. For instance, considering the words *bioinformatics* and *biology*, the Damerau-Levenshtein distance is 7, as this is the minimum number of operations necessary to convert one word into the other.

The longest common prefix (LCP) between the tokens determines the maximum length of the starting segment shared by two tokens. Applying this metric to the words *bioinformatics* and *biology* the longest common prefix yields a result of 3, corresponding to the length of the shared prefix *bio*.

The similarity score s between two tokens, t_1 and t_2, is calculated as the maximum of the two values:

$$s = \max\left(1 - \frac{DL_{(t_1,t_2)}}{\max[len(t_1), len(t_2)]}, \frac{LCP_{(t_1,t_2)}}{\max[len(t_1), len(t_2)]}\right) \tag{6}$$

In the example of *bioinformatics* and *biology*, the value of the similarity score is 0.5:

$$s = \max\left(1 - \frac{7}{14}, \frac{3}{14}\right)$$

2. Calculate the *soft* intersection count by considering the sum of similarities for the best-matching pairs of tokens between two sets, without repetition. In formula:

$$I_s = \sum_{(t_1,t_2)\in I} \mathrm{s}(t_1, t_2) \tag{7}$$

3. Compute the *soft* union defined as the total number of tokens minus the soft intersection count. In other words, the union in the context of the soft Jaccard score includes all unique tokens from both lists, adjusted for the soft intersections. This is essentially the sum of the lengths of both token lists minus the soft intersection count.

$$U_s = \mathrm{len}(T_1) + \mathrm{len}(T_2) - \mathrm{I} \tag{8}$$

Given the steps above, the Soft Jaccard Score is given by the ratio between soft intersection and soft union:

$$SJ(T_1, T_2) = \frac{I_s}{U_s} \tag{9}$$

Applying these steps to the example of T_1 = [bioinformatics] and T_1 = [biology], we have[8]:

$$\begin{aligned} I_s &= 0.5 \\ U_s &= 1 + 1 + 0.5 \\ SJ &= \frac{0.5}{1.5} = 0.333 \end{aligned}$$

For each degree title, we computed the value of the Soft Jaccard Score against all subject content areas included in each education field. For each education field, then, we retained the maximum value of this score. For example, in the case of bioinformatics, the output of this operation is the following:

Among all education fields, we classified the degree title in the education field with the highest Soft Jaccard score overall. In the case of *bioinformatics*, this degree title was classified as ICT as this is the field registering the highest Jaccard score.

[8]Noe that, in this example, the two sets T_1 and T_2 consist of one token each. Needless to say, the measure can be applied to comparing sets consisting of more than just one word. We keep this example for consistency with the discussion in the previous sections.

Table A6.5: Bioinformatics: ISCED field and area with largest SJ

Education field	Subject content area	Soft Jaccard Score
Agriculture	Forestry	0.400
Humanities	Ethics	0.647
Business	Typing	0.474
Education	Didactics	0.474
Engineering	Robotics	0.556
Health	Anatomy	0.400
ICT	**Informatics**	**0.867**
Natural Sciences	Geoinformatics	0.750
Services	Gymnastics	0.474
Social Sciences	Civics	0.556

6.5 Putting together

After applying the four methods described above, we collected the results and compared the classifications produced by each of them. For example, in the case of bioinformatics, we would obtain a vector like this:

Table A6.6: Predicted education field from various models

Title	Word2Vec (Google News)	GloVe (Wikipedia)	GloVe (Common Crawl)	Soft Jaccard
Bioinformatics	Natural sciences	Natural sciences	Natural sciences	ICT

Given the outcomes produced by the four methods described, we adopted a conservative approach with a focus on high precision. We proceeded as follows:

- If all four methods (the three word-embedding cosine similarity classifiers and the Soft Jaccard Score) unanimously predicted the same education field for a degree title, we accepted this prediction and classified the degree accordingly within the ISCED-F 2013 framework.
- Conversely, if at least one of the four methods provided a discordant prediction, diverging from the consensus of the others, we undertook a manual review of the degree title. This degree was then classified into one of the 11 ISCED education fields based on our evaluation. This assessment leveraged both the detailed descriptions provided by the ISCED-F 2013 documentation and supplementary information sourced from the internet.

For instance, in the case of *bioinformatics*, since the four methods produce discordant results, we classified this degree under the field of natural sciences after consulting various authoritative websites.

Our classification was limited to education titles equivalent to a BSc or higher, including MSc, PhD, and postgraduate diplomas. Consequently, no degree titles at the secondary or post-secondary education level were classified.

The table below presents the distribution of the 13,878 founders according to their field of education, detailing both the number and percentage of founders within STEM, Business, and Other categories. It is important to note that the sum of percentages does not reach 100 (nor does the sum of founders in each category equal the total number of founders) due to some individuals having degrees in multiple fields of education, such as an MSc in Engineering coupled with an MBA.

Table A6.7: Distribution of founders by education field

ISCED field	No. of founders	%
00 – Generic programs and qualifications	195	1.4%
01 – Education	139	1.0%
02 – Arts and humanities	1080	7.8%
03 – Social sciences, journalism and information	1675	12.1%
04 – Business, administration and law	4206	30.3%
05 – Natural sciences, mathematics and statistics	1350	9.7%
06 – Information and communication technologies	2961	21.3%
07 – Engineering, manufacturing and construction	3204	23.1%
08 – Agriculture, forestry, fisheries and veterinary	36	0.3%
09 – Health and welfare	339	2.4%
10 – Services	159	1.1%
STEM (05, 06, 07)	6677	48.1%
Business (04)	4206	30.3%
Others	3249	23.4%

Note: The sum of percentages does not equal 100 due to some individuals having degrees in multiple education fields (e.g. an MSc in engineering and an MBA, or BSc in engineering and MSc in computer science).

The majority of founders are graduates in STEM fields (48%), followed by Business (30%), and Other fields (23%). These figures are somewhat different from those reported in Retterath and Braun (2020). For the sample studied by these two authors, graduates in STEM are only 29% of all founders, whereas 39% graduated in business disciplines. In this respect, it is worth noting again that the samples studied in this paper and the one examined by Retterath and Braun (2020) are quite different. Whereas they consider only startups that have received multiple funding rounds from VCs, our sample comprises firms that have received one funding round from any type of investor (and that have not necessarily

received other funding).

Appendix 7 Classification of startup sector

Startups in the Crunchbase (CB) database are designated one or more *industry category* tags, referred to within the database as industry category groups. Our dataset features 46 unique tags, enumerated in Table A7.1. This system allows for a multifaceted classification of startups, where, for instance, a company could be tagged both under lending and investments and financial services, reflecting its diverse operational focus. In our analyzed sample of 6,867 startups, the distribution of tags per startup ranges from a minimum of 1 to a maximum of 14, with an average of 2.89 tags per company.

Table A7.1: CB Industry category groups

Code	Category	Code	Category
1	Administrative services	24	Information technology
2	Advertising	25	Internet services
3	Agriculture and farming	26	Lending and investments
4	Apps	27	Manufacturing
5	Artificial intelligence	28	Media and entertainment
6	Biotechnology	29	Messaging and telecommunications
7	Clothing and apparel	30	Mobile
8	Commerce and shopping	31	Music and audio
9	Community and lifestyle	32	Natural resources
10	Consumer electronics	33	Navigation and mapping
11	Consumer goods	34	Payments
12	Content and publishing	35	Platforms
13	Data and analytics	36	Privacy and security
14	Design	37	Professional services
15	Education	38	Real estate
16	Energy	39	Sales and marketing
17	Events	40	Science and engineering
18	Financial services	41	Software
19	Food and beverage	42	Sports
20	Gaming	43	Sustainability
21	Government and military	44	Transportation
22	Hardware	45	Travel and tourism
23	Health care	46	Video

To categorize each startup into a singular *sector*, we employed Latent Class Analysis (LCA). LCA is notably apt for handling data with categorical attributes, operating through an iterative, maximum likelihood estimation process. Initially, startups are randomly divided into a predetermined number of classes. Subsequent iterations reclassify these startups to enhance the model fit, continuing until convergence is achieved — indicated by a state where

no notable improvement can be made. The optimal classification, as determined by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), distributed our startups across 12 distinct clusters. These clusters are detailed in Table A7.2.

Table A7.2: Distribution of startups by sector of activity (LCA analysis)

Sector	Number of firms	Percentage (%)
Biotech & life sciences	298	4.3
Commerce	1837	26.8
Data analytics	413	6.0
Design & fashion	114	1.7
Fintech	485	7.1
Green tech & energy	232	3.4
Hardware	433	6.3
Internet services	195	2.8
Media & entertainment	834	12.1
Mobile apps	389	5.7
Sales & marketing	230	3.3
Software	1365	19.9
Not classified	42	0.6
Total	6867	100.0

The distribution of startups in our sample predominantly falls within three key sectors: (e-)Commerce, Software, and Media & Entertainment, which constitute 27%, 20%, and 12% of the total, respectively. Together, these fields represent approximately 59% of all startups in our analysis.

It is important to note that 42 startups could not be assigned to any of the 12 clusters due to the absence of industry tags in the CB database. These firms were subsequently excluded from our analysis, resulting in an adjusted sample size of 6,824 firms.

Appendix 8 Coding skills

Recognizing that formal education and training constitute only a portion of a person's knowledge base, we acknowledge the critical role of on-the-job training and experience in shaping a startup founder's background.

Initially, we sought to utilize the job titles listed on LinkedIn and Crunchbase as indicators of founders' professional experiences. Unfortunately, this method proved to be ineffective. The primary challenge lies in the non-standardized and highly personalized manner in which individuals report their job titles, leading to substantial variability in the representation of occupations. Furthermore, most job titles are overly generic, rendering them ineffective for distinguishing between *tech* and *business* experience.

Table A8.1: Top 20 job titles reported in LinkedIn

Job title	Frequency	Percentage on total
co-founder	1741	2.81
founder	1337	2.16
ceo	799	1.29
software engineer	630	1.02
intern	500	0.81
co founder	494	0.80
consultant	469	0.76
founder & ceo	428	0.69
cto	384	0.62
co-founder & ceo	349	0.56
project manager	341	0.55
software developer	339	0.55
director	321	0.52
research assistant	316	0.51
owner	289	0.47
founder and ceo	258	0.42
product manager	243	0.39
president	240	0.39
ceo & co-founder	223	0.36
web developer	223	0.36

For illustration, we compiled the top 20 job titles from the LinkedIn profiles of founders in our sample, as shown in Table A8.1. Notably, the most common titles pertain to roles as company founders. However, among these top 20 titles, only two—software engineer and software developer—are clearly associated with technical expertise. Excluding generic titles such as *founder*, *owner*, and *ceo* does little to alleviate this issue, as evidenced by the data in Table A8.2. The majority of the titles remain too ambiguous for effective classification.

Instead of continuing down the unproductive path of analyzing job titles, we pivoted to a different strategy by focusing on the *skills* section of founders' LinkedIn profiles. This section

Table A8.2: Top 20 job titles reported in LinkedIn, excluding founder, owner and ceo

Job title	Count	Percentage
software engineer	630	1.30
intern	500	1.03
consultant	469	0.97
cto	384	0.79
project manager	341	0.71
software developer	339	0.70
director	321	0.66
research assistant	316	0.65
product manager	243	0.50
president	240	0.50
web developer	223	0.46
partner	208	0.43
senior software engineer	203	0.42
managing director	193	0.40
advisor	192	0.40
associate	188	0.39
analyst	168	0.35
developer	160	0.33
researcher	159	0.33
board member	156	0.32

offers a significant advantage: it consists of a list of qualifications and abilities self-reported by individuals in a more standardized format than job titles. These skills serve as reliable indicators of a founder's experience domain. For instance, a founder skilled in ⟨C#, C++, Java, MySQL⟩ likely has a technical background, whereas someone with skills in ⟨Business Development, Business Strategy, Marketing Strategy⟩ probably has business expertise.

To operationalize this approach, we first compiled a comprehensive list of skills from all startup founders identifiable on LinkedIn, extending beyond those in our specific sample. We then filtered out uncommon skills, defined as those appearing fewer than 30 times across all analyzed profiles, resulting in a working pool of 2,878 distinct skills. For each pair of skills $[i, j]$, we calculated the Jaccard coefficient, defined as follows:

$$J_{ij} = \frac{a}{a + b + c}$$

where a is the number of times that skills $[i, j]$ co-occur in founders' resumes, b is the total number of times that skill i occurs without j, and c is the total number of times that skill j occurs without i

This metric quantifies the relatedness between two skills, reflecting the extent to which they co-occur in the resumes of founders. A higher frequency of co-occurrence indicates a stronger relationship between the skills in terms of shared competencies and abilities. The

Jaccard coefficient ranges from 0, indicating that skills i and j never appear together in any resume, to 1, signifying that skills i and j are always mentioned together in the resumes of founders.

Then, we constructed the adjacency matrix $\mathbf{W}$, whose cell w_{ij} contains the relatedness value (i.e., Jaccard coefficient) between skill i and skill j. Using this matrix, we applied a modularity-based cluster analysis. This type of clustering approach partitions a network (in our case the adjacency matrix $\mathbf{W}$) into distinct groups of vertices (also called modules or communities), such that connections within each group are dense but between groups are sparser (Newman 2006). Given a partition of the network, its quality can be assessed through the so-called modularity value, a quantity that ranges from -1 to $+1$ and measures the density of links within groups of vertices as compared with links between groups. For a weighted network, this quantity is given by the formula below where w_{ij} represents the weight (i.e., relatedness) of the edge between i and j, $k_i = \sum_j w_{ij}$ is the sum of the weights of the edges attached to vertex (i.e., skill) i, c_i is the group (or community) to which vertex i is assigned, the δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, and $m = \frac{1}{2}\sum_{ij} w_{ij}$. (Blondel et al. 2008; Newman 2004):

$$Q = \frac{1}{2m}\sum_{ij}\Big(w_{ij} - \frac{k_i k_j}{2m}\Big)\delta\Big(c_i c_j\Big)$$

Among the various algorithms suggested in the literature for optimizing modularity in network partitions, we adopted the approach proposed by Blondel et al. 2008. Specifically, we utilized the `community` method from the Python package `networkx` for our analysis. This process resulted in the partitioning of skills into 12 major clusters. Each cluster was labeled based on a detailed examination of the predominant skills within. Figure A8.1 showcases the network of skills, where vertices represent individual skills and edges denote their relatedness, quantified by the Jaccard coefficient. For clarity, only the edges representing the highest degrees of relatedness are depicted. Each skill cluster is identified with a unique label and visualized in a distinct color.

Subsequently, we associated each founder in our dataset with the relevant skill cluster, categorizing their skills accordingly. For instance, skills such as $\langle$\{C#: software, C++: software, Java: software, MySQL: software\}$\rangle$ were identified.

Because our purpose is to classify founders according to whether they have a *Tech* or a *Business* experience background, we considered as *Tech* all skills classified as software, engineering, life sciences, and cybersecurity. We considered as *Business* all skills classified in the clusters business, finance, and law.

In the final step of our analysis, we aimed to ascertain each founder's background as

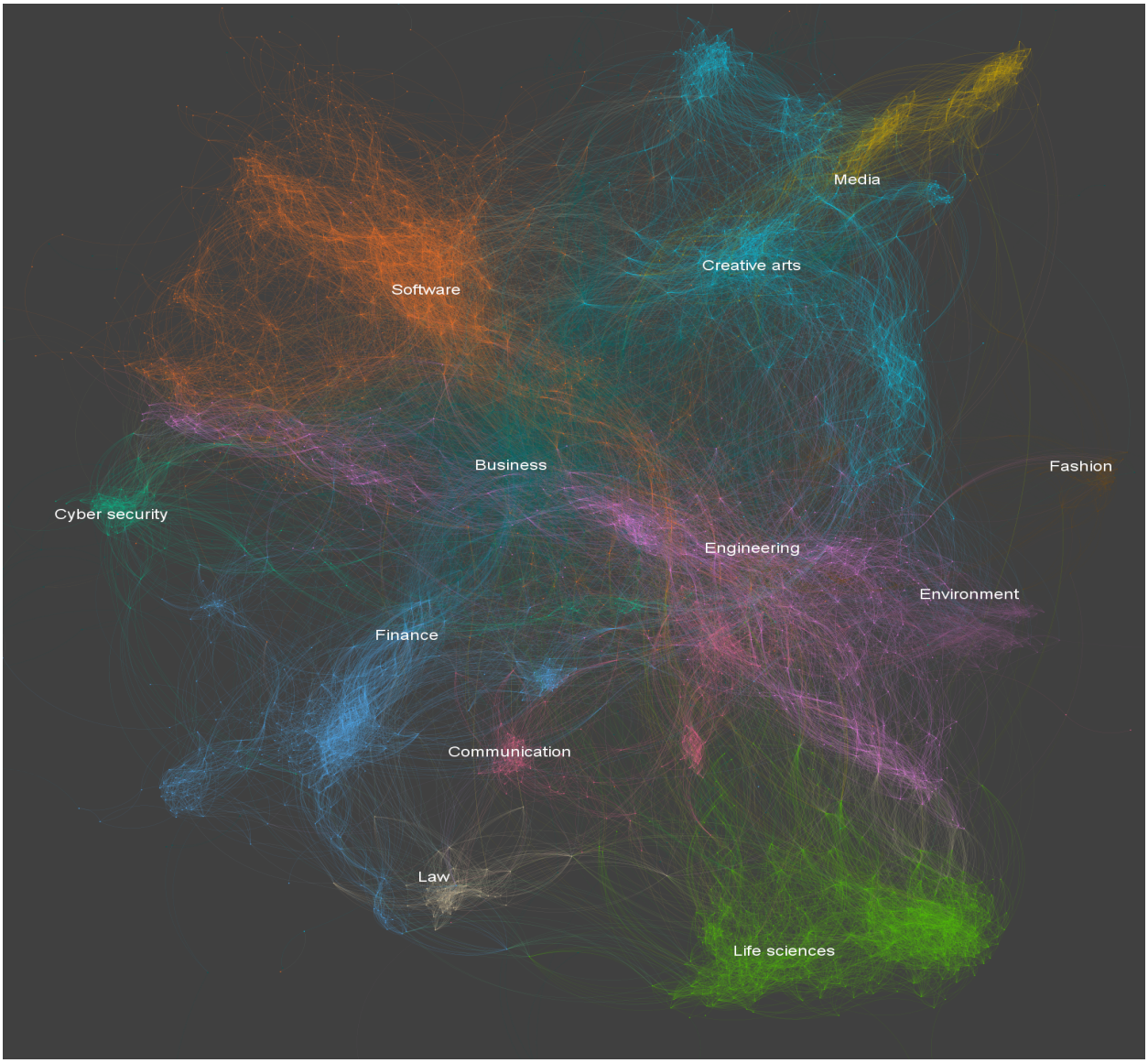


Figure A8.1: Community structure of skills

either *Tech*, *Business*, or other. This was achieved by calculating a relative specialization index for each founder, which allowed us to quantitatively differentiate their core areas of expertise based on the classified skills. The index is defined as follows:

$$Spec_{i,j} = \frac{\frac{n_{i,j}}{\sum_j n_{i,j}}}{\frac{\sum_i n_{i,j}}{\sum_i \sum_j n_{i,j}}}$$

where $n_{i,j}$ represents the number of skills in founder i's resume classified in experience field j (with $j = [Business, Tech, Other]$). The specialization index takes values greater than one for founder i when the share of his/her skills in field j is greater than the share that this field has in the overall population of founders. The field in which this specialization index takes the greatest value is assumed to represent the most distinctive area of background experience of a founder.

Using this method, we redefined the classification of startups. Specifically, we defined a startup as:

- *Pure tech startup* if at least one of its founders has a STEM education **or** has a background experience in *Tech* **and** none of its founders has either *business* education **or** a background experience in *Business*
- *Business startup* if at least one of its founders has a business education **or** has a background experience in *Business*

Table A8.3: Nature of startups: education and experience background

		Based on education		
		Business	Pure tech	Other
Based on experience background	Business	100.0	28.7	34.6
	Pure tech	0.0	71.3	14.9
	Other	0.0	0.0	50.5
	Total	100.0	100.0	100.0

Given our revised criteria for categorizing startups based on the founders' education *or* experience backgrounds, we anticipated an increase in the proportion of startups identified as *Business startups*. Table A8.3 validates this hypothesis. By our methodology, all startups previously classified as *business startups* due to their founders' educational background retain their classification when considering experience background. Conversely, 28.7% of startups initially categorized as *Pure tech* are reclassified as *Business* startups. This shift

occurs as a result of recognizing business-related skills in their founders' experience profiles, thereby broadening the definition of a *Business* startup to encompass both educational and experiential dimensions.

Appendix 9 Data validation

A potential concern in our study refers to the coverage of accelerated firms in Crunchbase. To the extent that (1) the coverage of accelerated firms in Crunchbase is incomplete and (2) the included firms are not a reasonably random sample, our analysis and results could be biased.

To mitigate these concerns, we carried out a validation exercise, triangulating CB data with that coming directly from the accelerators' websites. Specifically, we concentrated on two leading accelerators: Y Combinator and Techstars. From their respective websites, we gently scraped information on accelerated startups, selecting those up to the 2018 cohort (inclusive), yielding 1,747 startups for Y Combinator and 1,580 startups for Techstars. The lists of scraped startups are available in Stata format here for Y Combinator and here for Techstars.

Subsequently, we matched these startups with those in Crunchbase to assess (i) the extent of database coverage and (ii) the reasons for their inclusion (exclusion) from our sample. This matching task was intricate due to the potential name changes of startups and the presence of different identities and duplicates in both databases. Despite our diligence, minor errors might still exist. The results of this matching effort are available in Stata format here for Y Combinator and here for Techstars.

9.1 Y Combinator

Table A9.1 summarizes the evidence for Y Combinator. Of the 1,747 startups that went through acceleration at this accelerator between 2005 and 2018, we did not find any information on Crunchbase for only 24 companies (i.e., 1.4% of the total). For 32 startups (1.8%), we found the company in Crunchbase, but only in a release of the database post-2020. Since we utilized the 2020 release of Crunchbase for our analysis, these 32 companies were not included. Overall, the version of Crunchbase used for this paper covers approximately 96.8% of all startups accelerated at Y Combinator, indicating that the coverage of startups accelerated by Y Combinator is reasonably complete.

An additional 38 startups (i.e., 2.2% of the YC population) were excluded because no information on the first round of investment was available in Crunchbase. In other words, although these startups were listed in Crunchbase, the database lacked any details regarding the identity of the investor or the amount of the first round investment.

Focusing on the reasons why our final sample includes a relatively small number of Y Combinator firms, we considered the remaining 1,653 Y Combinator companies (i.e., $1,747 - 24 - 32 - 38$), which are represented in Crunchbase and for which information on

the first round of investment is available. From Table A9.1, it is evident that we excluded 497 startups (i.e., 30% of the 1,653 firms) from our sample, not due to a lack of information, but because their first round of investment did not involve Y Combinator. It is crucial to note that, since our sampling strategy was specifically aimed at startups whose *first round* investment was made by an accelerator, the exclusion of these companies from our sample does not reflect a deficiency in the database utilized or in our data gathering approach.

Table A9.1: Matching Y Combinator and Crunchbase: summary

Reason	**Number of firms**
Not found in CB	24
In CB, but in the new release of database (> 2020)	32
In CB, but no information on first-round investment	38
First round investor in CB is NOT Y Combinator	497
First round investor in CB is Y Combinator	1,156
All Y Combinator startups: cohorts ≤ 2018	1,747

Of the 1,156 startups that received a first-round investment from Y Combinator, our sample includes 448 companies. This constitutes 39% of all startups with a first-round investment from YC. To elucidate the reasons behind the exclusion of the remaining 61%, we carefully identified the factors leading to the removal of these companies from our final sample. The findings of this analysis are detailed in Table A9.2.

Table A9.2: Y Combinator: Reasons for sample selection

Index	**Number of firms**
Founded before 2004	11
First investment round after 2018	9
No information on founder names	21
No information on education of all founders	122
No information on company sector	1
First round investment greater than 150k	194
First round amount from YC missing in CB	350
Our final sample	448
Startups with a first investment round from YC	1,156

We excluded 11 startups because they were founded before 2004, adhering to our selection criterion that companies must be established between January 1, 2004, and January 1, 2018. Similarly, 9 companies were removed because their first investment round, as reported in Crunchbase (CB), occurred after 2018, contravening our criterion that the first investment

round must take place before December 31, 2018. This exclusion highlights potential inaccuracies in the recording of investment dates in Crunchbase, given our focus on YC cohorts from 2005 to 2018 (inclusive).

Additionally, 21 startups were removed due to the absence of founder names in Crunchbase and LinkedIn. Another 122 startups were excluded because, although founder names were available, we could not ascertain the educational backgrounds of all founders. One company was eliminated from our sample because it lacked information on its sector of activity.

The most significant reduction in our sample was due to financial criteria. We removed 194 startups for which the amount of the first investment round exceeded $\$150,000$ USD. Furthermore, we excluded 350 firms because Crunchbase failed to report the amount of the first investment round from YC. For these companies, while Crunchbase documented the investment date and the investor's identity (YC in this instance), the field for the raised amount was left blank. Opting against imputing an arbitrary amount, we decided to exclude these companies from our sample.

As previously mentioned, of the $1,156$ startups sourced from the YC website and identified in Crunchbase as having received a first investment round from YC, our final sample consists of 448 firms, approximately 39% of the total.

9.1.1 Y Combinator: Comparing included and excluded startups

To assess if the 448 startups included in our analysis constitute a reasonably random and unbiased subset of the $1,156$ startups that, according to Crunchbase (CB), underwent acceleration at Y Combinator, we conducted a series of tests. These tests were designed to compare the included and excluded startups across various dimensions.

Table A9.3: Y Combinator: Comparing included and dropped startups (t-tests)

	Mean		t-test	p-value
	Dropped	Included in sample		
Successful exit (IPO/Acquisition)	0.0969	0.1282	-1.6229	0.1046
Number of funding rounds	2.4736	2.3462	-1.1558	0.248
Total funding raised ('000 USD)	31224.1481	27604.5591	-0.2503	0.8024

Note: This table compares the average outcomes of startups that were included in the sample versus those that were dropped, with statistical significance assessed via t-tests. Successful exit is the share of startups that experienced IPO or acquisition. The difference is assessed via a z-test.

In Table A9.3, we evaluate the differences between the included and excluded startups in terms of the average number of funding rounds and the total amount raised. A t-test

Table A9.4: Y Combinator: Distribution of startups across sectors

	Dropped	Included in sample
Biotech	0.0441	0.0128
Commerce	0.2489	0.1581
Data analytics	0.0485	0.0527
Design and fashion	0.0154	0.0157
Fintech	0.0925	0.1054
Green energy	0.0242	0.0142
Hardware	0.0705	0.0741
Internet services	0.0463	0.0499
Media and entertainment	0.0947	0.1168
Mobile apps	0.0661	0.1182
Sales and marketing	0.0198	0.0271
Software	0.2291	0.2308
Total	1.0000	1.0000

reveals no significant differences between the two groups. Furthermore, we examined the proportion of startups in each sample that achieved a successful exit, either through an IPO or acquisition. A z-test confirms that the samples do not significantly differ in this aspect.

Table A9.4 presents a comparison of the industry distribution between the startups included in the sample and those excluded. Our analysis reveals a lower representation of biotech and commerce sectors and a higher representation of mobile apps within the included firms, as opposed to the excluded ones. Beyond these observations, there are no substantial differences between the two groups. A Mann-Whitney U test, yielding a statistic of 69.0 and a p-value of approximately 0.88, indicates that the distributions across industries between the two samples are not significantly different.

9.2 Techstars

In a similar analysis for Techstars, out of the $1,580$ firms scraped from the official Techstars website, we were unable to locate only 23 startups in Crunchbase (Table A9.5), representing less than 1.5% of the total. For 18 startups, although they were found in Crunchbase, the database lacked any information on investment rounds. Additionally, 553 startups were excluded from our sample because the first investment round reported in Crunchbase did not involve Techstars. Consequently, we were left with 986 startups that, according to Crunchbase, received their first investment round from Techstars.

For the 986 startups identified as having their first investment round with Techstars, we examined the factors leading to their exclusion from our final sample. We excluded 27 firms due to their foundation before 2004, reporting of the first investment round in Crunchbase

Table A9.5: Matching Techstars and Crunchbase: summary

Reason	**Number of firms**
Not found in CB	4
In CB, but in the new release of database (> 2020)	19
In CB, but no information on first-round investment	18
First round investor in CB is NOT Techstars	553
First round investor in CB is Techstars	986
All Techstars startups: cohorts ≤ 2018	1,580

(CB) after 2018, or the absence of information on the identity of the founders. Additionally, 94 firms were removed due to the lack of information on the educational background of all founders, and 117 firms were excluded because the amount of the first investment exceeded $150,000 USD. Similar to the situation with Y Combinator, the predominant factor leading to a significant reduction in our sample was the omission by Crunchbase of the investment amount in the first round for 518 startups. Consequently, our final sample comprises 230 startups, representing approximately 23% of the 986 firms that underwent a first investment round with Techstars.

Table A9.6: Techstars: Reasons for sample selection

Index	**Number of firms**
Founded before 2004	2
First investment round after 2018	1
No information on founder names	24
No information on education of all founders	94
First round investment greater than 150k	117
First round amount from TS missing in CB	518
Our final sample	230
Startups with a first investment round from TS	986

9.2.1 Techstars: Comparing included and excluded startups

To assess the representativeness of the 230 startups included in our analysis as a reasonably random and unbiased subset of the 986 firms identified in Crunchbase (CB) as having undergone acceleration at Techstars, we conducted a series of tests. These tests aimed to compare the included and excluded startups across various dimensions, ensuring the robustness and reliability of our findings.

Table A9.7 demonstrates that there are no statistically significant differences between the startups included in our sample and those excluded in terms of the average number of funding rounds, the average total funding raised, and the proportion of startups achieving a successful exit.

Regarding the sector distribution, Table A9.8 indicates that our analyzed sample has a smaller percentage of startups in commerce and a larger percentage in data analytics compared to the companies excluded. Nonetheless, a Mann-Whitney U test, with a statistic of 66.0 and a p-value of approximately 0.75, suggests that the differences in sector distribution between the two groups are not statistically significant.

Table A9.7: Techstars: Comparing included and dropped startups (t-tests)

	Mean		t-test	p-value
	Dropped	Included in sample		
Successful exit (IPO/Acquisition)	0.0958	0.0818	0.6789	0.4972
Number of funding rounds	2.9458	3.0985	1.0416	0.2982
Total funding raised ('000 USD)	4251.5179	5476.6597	1.2201	0.2229

Note: This table compares the average outcomes of startups that were included in the sample versus those that were dropped, with statistical significance assessed via t-tests. Successful exit is the share of startups that experienced IPO or acquisition. The difference is assessed via a z-test.

Table A9.8: Techstars: Distribution of startups across sectors

	Dropped	Included in sample
Biotech	0.0208	0.0121
Commerce	0.2417	0.1247
Data analytics	0.0917	0.1327
Design and fashion	0.0167	0.0268
Fintech	0.1083	0.1300
Green energy	0.0125	0.0255
Hardware	0.0708	0.1032
Internet services	0.0375	0.0550
Media and entertainment	0.0792	0.0804
Mobile apps	0.0750	0.0764
Sales and marketing	0.0500	0.0308
Software	0.1958	0.1957
Total	1.0000	1.0000

9.3 Conclusions from the validation exercise

We believe the results of this validation exercise compellingly demonstrate that Crunchbase may serve as a reliable information source regarding the coverage of accelerated startups. Of the 1,747 startups listed on Y Combinator's official website and the 1,580 from Techstars, more than 96% were found in Crunchbase. Furthermore, for 86% of the startups associated with Y Combinator and 92% of those with Techstars, Crunchbase effectively identifies not only the companies' existence but also their specific affiliations with these accelerators.

This validation exercise further underscores a critical aspect of our sampling methodology: a considerable number of firms (28% for Y Combinator and 35% for Techstars) were excluded from our analysis precisely because their involvement with an accelerator did not constitute their *first investment round*—a key criterion for inclusion in our study. This observation is vital when considering the suitability of accelerator websites as primary data sources for our sample. Such websites often lack the granularity needed to discern the sequence of investment rounds, making it challenging to ascertain whether an accelerator's contribution constituted the startup's initial funding. Therefore, a comprehensive database like Crunchbase (CB) becomes indispensable for identifying this specific aspect of startups' funding histories. This reliance on CB is not merely a preference but a methodological necessity, given our study's focus and the limitations of alternative data sources.

The primary reason for excluding the remaining companies is that Crunchbase fails to report the funding amount of the first investment round. Although it acknowledges the startup's acceleration by Y Combinator or Techstars, the specific investment amount by the accelerator is not disclosed. While it might be possible to *impute* these values, such imputation would inevitably be somewhat arbitrary and open to debate, leading us to omit these startups from our sample.

Given that the absence of reported investment amounts constitutes the main exclusion criterion, we argue that our sample is unlikely to be systematically biased. It seems reasonable to assume that Crunchbase's omission of this detail is unrelated to the most relevant characteristics of a company. Through our analysis, we endeavored to demonstrate negligible differences between the analyzed sample and the excluded startups. Although our findings are based solely on two accelerators, they suggest that our dataset is a reasonably representative sample of accelerated firms, allowing for some generalization of our results.

References

Blondel, Vincent D. et al. (2008). "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. arXiv: 0803.0476.

Mikolov, Tomás, Kai Chen, et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv: 1301.3781.

Mikolov, Tomás, Ilya Sutskever, et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3111–3119.

Newman, M. E. J. (2004). "Analysis of weighted networks". *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 70.5, p. 056131. arXiv: 0407503 [cond-mat].

— (2006). "Modularity and community structure in networks." *Proceedings of the National Academy of Sciences of the United States of America* 103.23, pp. 8577–82.

Retterath, Andre and Reiner Braun (2020). "Benchmarking Venture Capital Databases". *SSRN Electronic Journal*.

Roche, Maria P., Annamaria Conti, and Frank T. Rothaermel (2020). "Different founders, different venture outcomes: A comparative analysis of academic and non-academic startups". *Research Policy* 49.10, p. 104062.

Russ, Daniel E. et al. (2016). "Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies". *Occupational and Environmental Medicine* 73.6, pp. 417–424.