TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Entrepreneurial Finance 2

Prof. Dr. Reiner Braun

Essays on Machine Learning and the Value of Data in Venture Capital

Andre Retterath

Vollständiger Abdruck der von der Fakultät für Wirtschaftswissenschafter der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.) genehmigten Dissertation.

Vorsitzende:	Prof. Dr. Dr. Ann-Kristin Achleitner
Prüfer der Dissertation:	1. Prof. Dr. Reiner Braun
	2. Prof. Dr. Christoph Kaserer

Die Dissertation wurde am 18.09.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Wirtschaftswissenschaften am 15.11.2020 angenommen.

.

I never guess.
It is a capital mistake to theorize before one has data.
Insensibly one begins to twist facts to suit theories,
instead of theories to suit facts.

Sherlock Holmes, “A Study in Scarlett” (Arthur Conan Doyle)

Acknowledgements

The completion of this dissertation required a great deal of support from many people, and I was privileged to have received this support throughout the course of my doctoral studies.

First, I would like to sincerely thank my academic supervisor, Prof. Dr. Reiner Braun. Without your willingness to supervise me in parallel to my practical work as a venture capital investor, this dissertation would not have been possible. Thank you so much for always being open to unconventional research ideas and helping me to become a truly independent thinker. Through your openness and interest in cross-departmental research, you helped me to grow and combine my deep interests in computer science and finance. Your great ambitions, together with your extremely positive and fun way of leading our team, impressed me from day one. Similarly, I owe deep gratitude to the Earlybird Venture Capital family and specifically to one of the founding partners, my mentor and practical supervisor, Dr. Hendrik Brandis. You have been a key driver and motivator for this dissertation, and I am extremely thankful for being able to work side-by-side with and continuously learn from you. Your extensive practical experience and thought-provoking impulses helped me to raise this academic work to the next level. Furthermore, I thank Prof. Dr. Stelios Kavadias, who took a keen interest in my research and was always open for discussion. Through your critical questioning and constructive criticism, you inspired me to think about details and reiterate until perfection. Your sharpness, professional competence and personal support motivated and helped me to not lose sight of my objectives.

Special thanks go to my parents, Elke and Gerd Retterath, who always believe in me. I also thank my sister and her husband, Katrin and Michael Henße, for keeping me grounded and regularly reminding me of what really matters in life. Lastly, I heartily thank my wife, my better half and eternal cheerleader, Lisa Retterath, who always motivates me to never give up. Thank you so much for continuously covering my back and for supporting my innumerable night and weekend shifts. You are an integral part not only of my personal life but also of my professional development and, more specifically, of this dissertation.

Table of Contents - Overview

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AC	Accuracy
AI	Artificial Intelligence
AL	Angellist
BVD	Bureau van Dijk
CB	Crunchbase
CI	CBInsights
CH	Companies House
CVC	Corporate Venture Capital
DL	Deep Learning Models
DT	Decision Trees
DR	Dealroom
e.g.	exempli gratia
et al.	et alii
etc.	et cetera
FN	False Negative
FP	False Positive
fVCs	formal Venture Capitalists
GL	Generalized Linear Models
GP	General Partner
i.e.	id est
iVCs	informal Venture Capitalists
IRR	Internal Rate of Return
LP	Limited Partner
LR	Logistic Regressions

ML	Machine Learning
NB	Naive Bayes
NLU	Natural Language Understanding
P.A.	per annum
PB	Pitchbook
PE	Private Equity
PR	Prequin
RE	Recall
RF	Random Forests
SC	Selection Criteria
STEM	Science, Technology, Engineering, Math
TN	True Negative
TP	True Positive
TR	Tracxn
US	United States
VC	Venture Capital
VCs	Venture Capitalists
VS	VentureSource
XG	Gradient Boosted Trees or XGBoost Agorithm

1 Introduction

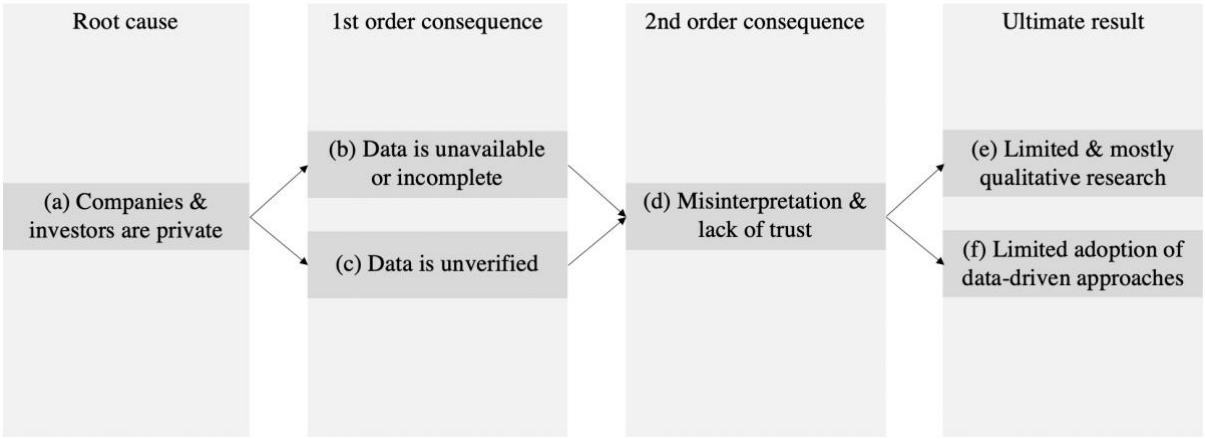
1.1 Motivation and research topic

Venture capitalists (VCs), who fall into the class of individuals and institutions that manage private capital assets, and private companies share a common denominator: the majority of their data are kept private. This leads to two closely related first-order consequences: First, information about VCs and their portfolio companies is frequently unavailable or incomplete, as neither entrepreneurs nor their investors are appropriately incentivized to share comprehensive and accurate information (Brealey, Leland, and Pyle, 1977; Ramakrishnan and Thakor, 1984). While entrepreneurs only selectively share information that supports their equity stories and intended market perceptions, VCs generally do not wish to share private information, as such information represents their competitive advantage and may expose their portfolio strategies or the performance of their funds. Hence, datasets are scarce and frequently proprietary, which prevents academics from empirically approaching novel research questions and replicating previous studies (Kaplan and Lerner, 2016). Second, most available information is unverified. It is unclear to what degree currently available data – which are often collected and distributed by commercial data platforms such as Crunchbase (CB), Pitchbook (PB) or CBInsights (CI) – are flawed. Due to the private nature of the represented companies and investors, data aggregators are unable to exhaustively verify the available information and thus frequently distribute unverified datasets (Kaplan and Lerner, 2016; Kaplan, Strömberg, and Sensoy, 2002; Maats, Metrick, Yasuda, Hinkes, and Vershovski, 2011). As a result of sparse, incomplete and unverified data, academics and VCs have difficulties in accurately interpreting the available information. They frequently draw opposing conclusions and misinterpretations based on different datasets, which leads to a lack of trust as the single most important second-order consequence. These factors ultimately result in two

major limitations: First, academic research on private companies and VCs has been heavily constrained and is largely qualitative in nature. As a consequence, important questions in the respective fields of study remain unanswered. Second, VC practitioners are skeptical of data-driven approaches and thus rarely adopt them to scale their operations (Arroyo, Corea, Jimenez-Diaz, and Recio-Garcia, 2019; Schmidt, 2019). This reluctance, in turn, explains why the VC investment process is still largely manual and subjective in nature. The way VCs operate has not materially changed since the inception of the asset class in the 1940s, whereas in other asset classes, such as hedge funds, operations have matured and started to scale by becoming heavily objective, quantitative and data-driven. Figure 1.1 summarizes the described cause-effect chain.

Figure 1.1: Cause-effect chain of private company data

This figure presents the cause-effect chain from the root cause of (a) companies and investors being private through to the first- and second-order consequences and to the ultimate results thereof. In summary, the first-order consequences are that (b) private company data are frequently unavailable or incomplete and that (c) available data are frequently unverified. This results in the second-order consequence that (d) data are frequently misinterpreted, and people do not trust the respective analyses. Ultimately, this leads to (e) limited and mostly qualitative research and (f) limited adoption of data-driven approaches within the VC industry.



This dissertation explores innovative approaches to solving the first- and second-order consequences of the lack of private company and investor information. By doing so, I seek to allow (e) academics to empirically approach unanswered research questions and (f) enable VC practitioners to scale their investment process through the adoption of data-driven methods. In summary, my goal is to identify new approaches to private data collection and data verification and to accelerate the adoption of data-driven applications within the VC investment process.

1.2 Development of research questions

This section starts with a brief overview of the general VC literature with a specific emphasis on the private data cause-effect chain as displayed in Figure 1.1. I seek to understand in detail how (b) unavailable, incomplete and (c) unverified information leads to (d) flawed analyses and misinterpretations and thus a lack of trust. Focusing on (e) an exemplary literature stream that is characterized by limited replicability and continuously opposing perspectives, I summarize its methodological shortcomings and highlight the potential for (b) novel data collection approaches to help (d) resolve this controversial debate and create the necessary trust. Moreover, I describe how (c) a quantitative benchmarking and verification of existing VC databases might help researchers to (d) interpret their results more accurately. Based on the assumption of (d) increased trust in such information, I further explore how (f) data-driven approaches possess the potential to scale the traditional, mostly manual and subjective VC investment process. Finally, this section outlines the structure of the remainder of the dissertation.

Pitchbook Research (2020) assumes that the private capital asset class controls around 98% of the U.S. economy by number of companies and more than 25% by amount of capital. After private equity (PE) firms that invest in more mature private companies or in publicly listed companies as part of "going private," VCs are the second subset of the private capital asset class (Wright and Robbie, 1998). The term VCs refers to (mostly private) investors who provide equity financing to (often newly formed) private companies in exchange for minority shareholdings (Achleitner, 2001). VCs can be firms that invest other people's money, so-called "formal VCs" (fVCs), or individual investors such as business angels who invest their private wealth, so-called "informal VCs" (iVCs). Section 2.1 describes the differences in more detail. Although the overall group of VCs accounts for a minority share of the private capital asset class with respect to the capital invested, its relative importance has been steadily increasing

(Gompers and Lerner, 2001; Gorman and Sahlman, 1989; Kaplan et al., 2005; Kortum and Lerner, 2001; Pitchbook, 2019; Pitchbook and NVCA, 2019). Despite this trend, Wright and Robbie (1998) already noted more than two decades ago that the academic literature was lagging behind the development of the VC industry. The authors' comparison of traditional corporate finance theory with the VC literature revealed that this divergence is likely rooted in one major difference: VC as a form of private capital relies on private information that is "*widespread and difficult to reveal,"* whereas, for traditional corporate finance, *"private information is rare and provisioning of public information is mandatory*." While the discrepancy between the importance of the VC asset class and the available research has become more pronounced, the root cause has not changed.

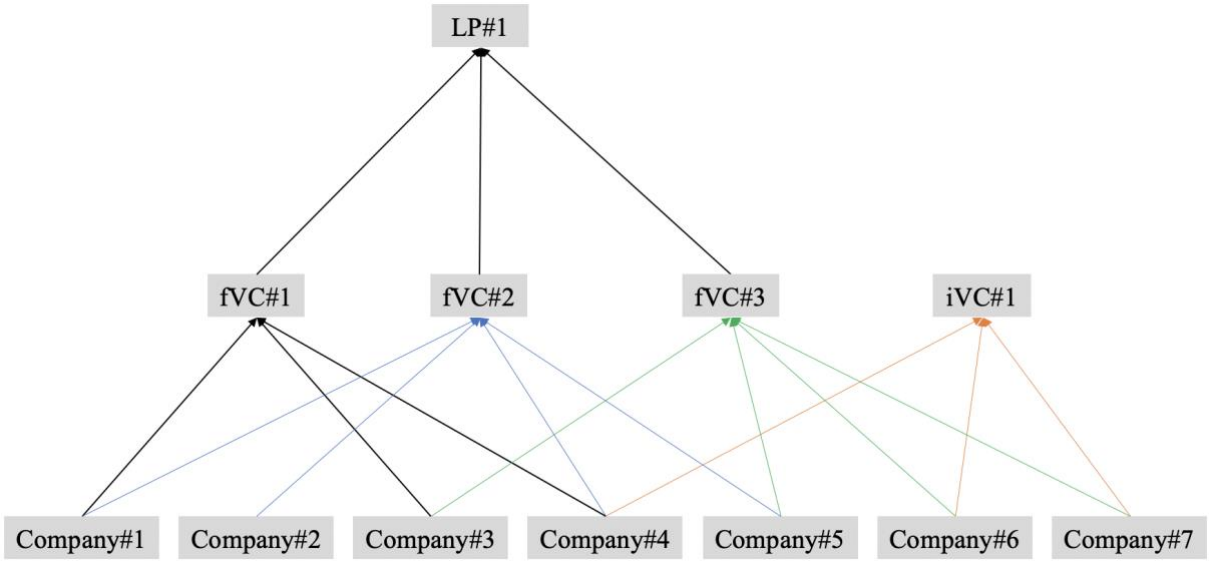
1.2.1 Data collection

To better understand why information on private companies and private VCs is "*widespread and difficult to reveal*," I disentangle and describe the information flow within the ecosystem below. On the most basic level, three types of players exist: (1) private companies; (2) VCs, which can be split into fVCs and iVCs; and (3) limited partners (LPs). fVCs collect cash commitments from various LPs; these sums are subsequently invested in multiple private companies, which constitute the portfolios of the fVCs. Consequently, fVCs serve as the link between their LPs and portfolio companies (Sahlman, 1990). Together with iVCs that invest their private wealth and thus do not have an upstream connection to LPs, the VC group is the focus of this dissertation. Although unidirectional reporting agreements and information rights exist between not only portfolio companies and their VCs but also between fVCs and their LPs, the information asymmetries increase with every additional intermediary. Such asymmetries arise as information is condensed and summarized at each level. A distinct research stream concerning the principle-agent theory investigates these information asymmetries, among other topics (Brealey et al., 1977; Jensen and Meckling, 1976). Comparing the available information

along the two dimensions of *sample size* and *level of detail* across private companies, VCs and LPs leads to the following picture: Private companies have information about a minimum number of companies (i.e., only about themselves), but with a maximum level of detail (deep). VCs have information about a medium number of companies with a medium level of detail. Finally, LPs have access to a maximum number of companies, but with a minimum level of detail (shallow). In summary, the three different types of players within the private capital ecosystem have access to different kind of information. Figure 1.2 summarizes the information flows between the respective groups.

Figure 1.2: Information flow among portfolio companies, VCs and LP

This figure presents the information flow among the portfolio companies on the bottom level, the VCs (which are split into fVCs and iVCs) on the mid-level and the LP (as an investor in fVCs) on the top level. The information flow is unidirectional from the bottom level to the top level, which means that the LP has information about different fVCs and their portfolio companies, whereas neither the fVCs nor their portfolio companies have information about the LP. Similarly, the fVCs and iVCs have information about the portfolio companies but not vice versa. It is important to note that the information is always condensed from the lower to the upper level (i.e., LPs have high-level summaries, whereas fVCs and iVCs have more granular information about the portfolio companies).



Due to the fact that all described players are private and thus have neither an obligation nor an appropriate incentive to share information with outsiders (exceptions include corporate VCs (CVCs) or publicly listed LPs), data are locked within these closed ecosystems (Freear, Sohl, and Wetzel Jr, 1994; Kaplan and Lerner, 2016; Kaplan et al., 2002; C. M. Mason and

Harrison, 2002; Ramakrishnan and Thakor, 1984). As a result, empirical research on private companies and VCs has been heavily constrained. In addition, VC practitioners encounter difficulties in actively identifying suitable investment targets, as they depend on the entrepreneurs providing the required information.

To mitigate the issue in academic contexts, scholars have explored a variety of creative data collection approaches. I cluster these approaches into three groups based on the channel of information collection ("entrepreneurs directly," "VCs directly" or "LPs directly") and distinguish them based on the resulting sample size, level of detail, typical type of research conducted on the basis of such datasets and freedom of sharing the resulting datasets.

Entrepreneurs directly. Scholars such as Breugst, Domurath, Patzelt, and Klaukien (2012), Domurath and Patzelt (2019), Rosenstein, Bruno, Bygrave, and Taylor (1993) and Segal, Borgia, and Schoenfeld (2005) have reached out to entrepreneurs directly to collect information about them or their companies. While entrepreneurs' low level of willingness to share complete and accurate information with external parties frequently reduces the conversion rate of such requests (Brealey, Leland, and Pyle, 1977), evidence shows that reaching out to befriended or otherwise closely connected contacts increases the resulting sample size (Gompers, Gornall, Kaplan, and Strebulaev, 2020). On the one hand, the process is highly manual and costly, which limits the number of companies willing to share information to a minimum. On the other hand, manual interaction makes it possible to achieve a maximum level of detail and obtain individual consent to freely share the resulting datasets. As a consequence of small datasets with a high degree of detail, the resulting research is mostly qualitative and explorative in its nature. It mainly seeks to understand the "why" and "how" on a case-by-case basis.

VCs directly. To overcome the limitations associated with the small sample size, researchers such as Gompers, Gornall, Kaplan, and Strebulaev (2020), Petty and Gruber (2011), Norton and Tenenbaum (1993) and Sandberg and Hofer (1987) have taken the "entrepreneurs directly" approach to the next level and reached out to "VCs directly" (both fVCs and iVCs) to collect

information about them and their portfolio companies. VCs act as a multiplier, as they centrally collect data about all of their portfolio companies. As a result, the company sample size increases from a minimum to a medium level, whereas the respective level of detail decreases from the maximum to a medium level due to the abovementioned condensing. Research based on such large-scale datasets is generally more quantitative in nature and addresses questions such as "how many" or to "what degree." While the "VCs directly" approach seems promising in terms of collecting large-scale portfolio company data and empirically addressing a variety of quantitative research questions, it has also been applied to collect information on VCs themselves. When used for this purpose, the approach features the same benefits and shortcomings as that of approaching entrepreneurs directly, only on the VC level. Due to the limited VC sample size, the resulting research is similarly qualitative and explorative in nature.

LPs directly. To further increase the portfolio company sample size and to overcome the limited VC sample size, researchers such as Maats et al. (2011) or Prencipe (2017) have raised the bar once more by approaching "LPs directly." While VCs act as a multiplier with respect to their downstream portfolio companies, LPs act as a multiplier with respect to their downstream VCs. In Braun, Jenkinson, and Stoff (2017)'s words, "*Funds, after all, are simply a legal wrapper around a sequence of underlying investments*." Therefore, the "LPs directly" approach leverages two multipliers with respect to portfolio companies, which further increases the company sample size from an intermediate to a maximum level and decreases the level of detail from intermediate to minimum. The same logic as for "VCs directly" with respect to their downstream portfolio companies applies for "LPs directly" with respect to their downstream VCs. The sample size increases from a minimum level to an intermediate level, and the level of detail decreases from the maximum to an intermediate level. Due to the fact that the resulting portfolio company and VC datasets are highly sensitive, sharing them is mostly prohibited.

The three approaches follow a pyramid logic in line with Figure 1.2, with "entrepreneurs directly" on the lowest level, "VCs directly" on the mid-level and "LPs directly" on the top

level. Due to the upstream reporting requirements, portfolio company and VC sample sizes are maximized when collecting information at the top (i.e., through "LPs directly"), whereas information is most detailed when it is collected from the relevant players on the mid or bottom levels directly (i.e., either through "VCs directly" in case the VC is of interest or "entrepreneurs directly" should the company or founders be of interest). I assume that the number of multipliers involved, the sensitivity of the data and the difficulty of collecting information are correlated (i.e., it should be easier to collect information from entrepreneurs whose access is limited to their own companies versus LPs who have access to information about multiple fVCs and their downstream portfolio companies). Matching existing academic studies with one of the three approaches and classifying the type of research as qualitative or quantitative leads to the identification of a clear pattern. Qualitative research questions such as "why" and "how" require detailed information but smaller sample sizes and thus rely on "entrepreneurs directly" for portfolio company-focused studies and "VCs directly" for VC-focused studies. These kinds of studies represent the majority of VC- and private company-focused literature. Quantitative research questions such as "how many" or "to what degree" require larger sample sizes but less detail and therefore rely on "LPs directly" for portfolio company- or VC-focused studies and "VCs directly" for portfolio company-focused ones. These kinds of studies are less common because the required datasets are, due to their sensitivity, more difficult to collect and cannot be freely shared for replication. Table 1.1 displays the described characteristics.

In summary, the three commonly applied data collection methods described above are either manual, very costly and result in smaller sample sizes or result in proprietary datasets that cannot be freely shared. To overcome these shortcomings, Section 2.1 addresses the following question:

(1a) How can we collect a large-scale private company dataset that can be freely shared and allows us to resolve highly relevant but unanswered research questions within the VC community?

Table 1.1: Characteristics of different data collection approaches

Overview of "Entrepreneurs directly", "VCs directly" and "LPs directly" approaches distinguished across the dimensions of sample size, level of detail, common research type and freedom to share the resulting datasets. As "VCs directly" and "LPs directly" can be applied to collect data on two or three different parties, respectively, these dimensions are further split into company, VC and LP to reflect the target group for which the information is collected.

	Entrepreneurs directly	VCs directly	LPs directly
Sample size			
Company	small	medium	large
VC	-	small	medium
LP	-	-	small
Level of detail			
Company	deep	medium	shallow
VC	-	deep	medium
LP	-	-	deep
Research type			
Company	qualitative	quantitative	quantitative
VC	-	qualitative	quantitative
LP	-	-	qualitative
Sharability			
Company	high	medium	low
VC	-	high	medium
LP	-	-	high

With respect to an unresolved research question, the issue of portfolio diversification versus specialization seems to be a suitable application for our novel data collection method, as this debate is probably one of the most controversial topics within the VC community (Bygrave, 1988; Cressy, Malipiero, and Munari, 2014; Cressy, Munari, and Malipiero, 2007; P. Gompers, Kovner, and Lerner, 2009; Gorman and Sahlman, 1989; Humphery-Jenner, 2013; Jackson III, Bates, and Bradford, 2012; Knill, 2009; Matusik and Fitza, 2012; Norton and Tenenbaum, 1993; Sahlman, 1990). The VC portfolio strategy research stream exemplifies how different proprietary datasets lead to varying conclusions while also preventing replication. As of the time of writing, there seems to be neither an agreement for fVCs (Buchner, Mohamed, and Schwienbacher, 2017) nor has this question been translated to iVCs (Antretter, Sirén, Grichnik, and Wincent, 2018; Bonini and Capizzi, 2019; Bonini, Capizzi, Valletta, and Zocchi, 2018; Freear et al., 1994). In line with the literature and the above-described fact that iVCs have no LPs, we assume that lack of research on iVCs' diversification strategies is due to the lack of large-scale datasets that can be freely shared. Hence, after exploring new avenues to efficiently

collect a comprehensive private company dataset in (*1a*), we leverage the resulting dataset to reverse-engineer VC portfolios (i.e., fVCs but more importantly iVCs) and empirically answer the following research question:

(1b) To what extent do existing research findings on the relationship between fVCs' portfolio diversification or specialization strategies and their returns translate to iVCs?

Note that Essay 1 is framed around research question (1b) rather than the methodological contribution of exploring a novel data collection approach (1a) so that it qualifies for submission to an entrepreneurial finance journal. In summary, Section 2.1 focuses on the end-to-end cause-effect chain ranging from a) to e) in Figure 1.1, with its major focus being on data collection.

1.2.2 Data verification

As the limitations associated with locked-in private company and VC information have been widely discussed, service providers identified a promising business opportunity and started to serve the gap between the increasing need for private data and the limited availability thereof. Such service providers collect private company and VC information once and then distribute it to an unlimited number of clients. While the unrestricted availability of these datasets ensures research replicability, it is unclear how these providers collect their data and whether these data are verified. Kaplan and Lerner (2016) and Kaplan et al. (2002), as well as Maats et al. (2011), have revealed that even some of the most established VC databases are incomplete and partially unverified. Moreover, Kaplan and Lerner (2016) noted with respect to the growing number of commercial data providers that "*While many of these newer databases are promising, they have not gotten the kind of scrutiny that VentureSource (VS) and VentureXpert have. Thus, their ability to support academic research is still to be fully determined.*" Despite this fact, an increasing number of academics and practitioners rely on such "newer databases" for lack of a better alternative (e.g., Achleitner, Braun, Behrens, and

Lange, 2019; Alexy, Block, Sandner, and Ter Wal, 2012; Bonini and Capizzi, 2019; Croce, Guerini, and Ughetto, 2018; D. E. de Lange, 2019; Ter Wal, Alexy, Block, and Sandner, 2016; Thies, Huber, Bock, Benlian, and Kraus, 2019; Winkler, Rieger, and Engelen, 2019). Clearly, accurate interpretation of the data contained in such databases requires a detailed understanding of potentially incomplete and biased information. Therefore, Section 2.2 addresses the following question:

(2) How well do the most prominent VC databases reflect actual information?

By approaching this question, we address (a), (c) and (d) of the cause-effect chain depicted in Figure 1.1 and seek to showcase a suitable data verification approach that can be replicated in order to scrutinize any kind of private company or VC dataset. I assume that our work will increase overall trust and the consistency of analyses based on the investigated databases.

1.2.3 Data application

In addition to academic researchers, VC practitioners are also highly interested in large-scale private company information that is structured and verified, as it has always been central in their investment process (Gompers, Gornall, Kaplan, and Strebulaev, 2020; Sahlman, 1990; Zacharakis and Meyer, 1998). In recent years, a number of data-driven approaches have been explored on the basis of such datasets, ranging from manual scorecards to automated machine learning (ML) models. The majority of these tools focus on the investment screening and selection process and aim to eliminate subjectivity to ultimately allow VCs allocating their limited resources more effectively (Arroyo et al., 2019; Catalini, Foster, and Nanda, 2018; Ghassemi, Song, and Alhanai, 2020; Krishna, Agrawal, and Choudhary, 2016). However, despite the increasing availability of such solutions, Schmidt (2019) found that "*In order to implement AI [in the VC investment process], organizational behaviors have to be changed slowly. The more often investment professionals are outperformed by the algorithm, the more the trust in the algorithm increases*." The interviews I conducted with 63 VCs, which are

presented in Section 2.3, confirm that 95% of VCs refrain from adopting such tools because they are concerned that a decline in screening performance might occur when compared to their status quo (i.e., investment screening conducted by human investment professionals). To solve this issue and create the trust necessary for VCs to adopt such data-driven tools, Section 2.3 considers the following research question:

(3) How does the performance of ML-based investment screening tools compare to the screening performance of VC investment professionals?

By conducting a comprehensive benchmarking study, I seek to provide the missing building block that can create the necessary trust in data-driven approaches. By doing so, I hope to encourage VCs to become less hesitant to adopt ML-based investment screening tools and start to integrate them into their existing workflows. Ultimately, I hope that the use of such tools will eliminate subjectivity in the screening process and help the VCs to allocate their resources more effectively.

1.3 Methods

As this dissertation comprises three distinct studies, the different contexts and research goals required the application of a diverse range of methodologies. These methodologies can be categorized into supervised ML algorithms and traditional regressions. Each of the following three paragraphs describes the specific methodological approaches applied in the data collection study, the data verification study and the data application study, respectively.

In Essay 1, we collected a novel dataset of 3,328 companies based in Cambridge, UK, which have cumulatively received 14,575 investments from 12,588 investors. We relied on Companies House (CH) and Bureau van Dijk (BVD) as the foundation of our dataset but automatically cleaned and verified the information using a rule-based data mining technique (Han, Pei, and Kamber, 2011). More specifically, we collected some "truth-level" data and matched them with the information in our dataset. Whenever we edited a datapoint, a new rule

was created through supervised ML. The resulting ruleset was automatically executed to edit and verify the existing dataset at scale. Subsequently, we ran an ordinary least squares (OLS) regression to answer our research question (1b). Moreover, we conducted a Poisson regression to test for robustness.

In Essay 2, we approached 10 European VC partnerships with which we have close relationships and collected information on 339 actual VC financing rounds from 396 investors in 108 different private (mostly European software) companies. Subsequently, we compared these "truth-level" data with their representation in the eight most prominent databases by focusing on their descriptive statistics. Additionally, we ran logistic regressions to understand potential biases and help researchers to interpret their results based on the investigated databases more accurately.

In Essay 3, I followed Arroyo et al. (2019) and Ghassemi et al. (2020) and trained a variety of supervised ML algorithms, including decision trees, deep learning models, generalized linear models, gradient boosted trees, logistic regressions, naive Bayes models and random forests, to predict venture success. In line with Arroyo et al. (2019), I used CB as the foundation of my dataset but complemented and verified specific variables with PB and LinkedIn information. As a result, my dataset comprises 77,279 European software companies that were founded after January 1st, 2010. I compared the performance of the trained ML algorithms based on a confusion matrix and selected the best performing one with regard to overall accuracy and recall. Subsequently, I compared the screening performance of the best performing algorithm (i.e., gradient boosted trees (XG)) with the performance of 111 human investment professionals in the same way. I collected the predictions of the investment professionals via an online survey. Ten different companies were presented to the investment professionals in the form of one-pagers, and I asked them to select exactly five companies for further analysis. As I provided them with historic input information, I was able to compare their predictions with the actual results and thus calculate their performance in a confusion matrix.

In summary, I showcase how a variety of ML algorithms and traditional regressions can be leveraged to collect, verify and apply private company data in VC contexts at scale. Using these approaches, scholars might be able to empirically address thus far unanswered research questions. Moreover, VC practitioners might apply these methods to achieve a more objective and scalable investment process.

1.4 Research results and contributions

Overall, this dissertation provides three major contributions that may help academics and practitioners to further gauge and unlock the potential of private data in VC. I seek to remove the persistent data barriers in academic VC research and accelerate the adoption of ML within the VC investment process. During my doctoral journey, I collected valuable insights as byproducts, which are described in detail in the individual studies. The main results, however, are summarized below.

First, we explore a scalable private company data collection approach that overcomes common sample size limitations and allows us to resolve a thus far unanswered research question. In light of the question whether portfolio diversification or specialization is more successful for fVCs and iVCs, we find a) that iVCs benefit less from industry and stage diversification than fVCs, b) that the properties of the investment strategies undertaken by investors (iVCs in particular) shed more light on the exact benefits of diversification and c) that the resulting returns are – independent of investor types – highly skewed (i.e., returns are distributed according to a power-law, and they depend on one or a few home run investments per portfolio).

Second, we pursue a comprehensive data verification approach by collecting actual contracts and investment documentation and comparing it with its characterisation in the eight most relevant VC databases. While the major driver of Essay 2 was showcasing a replicable data verification process, our benchmarking results help researchers and VC practitioners to

better understand the coverage and quality of their datasets and thus interpret their research results more accurately. More specifically, our analysis reveals that VS, PB and CB have the best coverage and are the most accurate databases across the dimensions of general company, founders and funding information. A combined dataset with the best possible coverage would consist of general company information from VS, founder information from PB and funding information from a combination of CB, PB and VS. With respect to the application of ML and data-driven investment screening approaches, the results of Essay 2 should help to increase the representativeness of the training data and remove potential biases from algorithms.

Third, I conducted several interviews with VCs to understand the adoption of ML and data-driven screening approaches within their investment processes. I find that VCs are hesitant to adopt such novel tools, mainly due to a lack of trust in the underlying data quality and the absence of a comprehensive benchmarking study that compares the performance of such tools with that of the VC status quo (i.e., human investment professionals). As Essay 2 addresses the data quality issue in detail, I assume that the performance benchmarking conducted in Essay 3 provides the missing building block required to create the necessary trust and accelerate the adoption of ML and data-driven screening tools. My comparison shows that the XG classification algorithm performs relatively 25% better than the *median* VC and 29% better than the *average* VC in screening and selecting European early-stage software companies. Similarly to Essay 1, this study also resulted in a valuable byproduct: I present a comprehensive characteristic-specific performance analysis for VC investment professionals that shows that a) institutional VCs perform better in the generic screen stage than CVCs; b) after approximately a decade of VC experience, there exists a negative correlation between experience and screening performance; c) there exists a positive correlation between the VCs' highest level of education and their screening performance; and that d) science, technology, engineering and math (STEM) graduates perform better than business graduates or graduates with other degrees. In summary, the best-performing venture capitalist for the generic screening of European early-

stage software companies would have the following profile: an investment professional with a Ph.D. in STEM and less than 10 years' VC experience who works for an institutional VC firm with more than €500 million in assets under management.

To conclude, this dissertation makes important contributions to the VC literature by exploring innovative private data collection and verification approaches and by providing evidence for the superior performance of ML-based investment screening tools. Moreover, it provides a number of valuable insights in the areas of VC portfolio strategy, return distribution and investment screening performance dependencies for VC investment professionals.

1.5 Dissertation structure and overview

The remainder of this dissertation is structured as follows: Chapter 2 consists of three self-contained essays across three different sections. Each section represents an individual academic paper that addresses a standalone research question. Section 2.1 explores the efficient and scalable collection of private company data that can be freely shared and leveraged to resolve highly relevant but yet unanswered research questions within the VC community. Based on the resulting dataset, we seek to answer the highly controversial question of whether portfolio diversification is more successful than portfolio specialization for fVCs and translate this question to the context of iVCs. Section 2.2 scrutinizes the most prominent private company and VC databases and aims to determine how well they reflect actual data. Here, we showcase a structured data verification approach that is intended to help researchers and practitioners interpret their results more accurately. Both essays seek to increase the availability and quality of private company and VC data and, as a consequence, increase the overall trust not only within the research community but also among VC practitioners. Subsequently, Section 2.3 explores a variety of data-driven approaches that aim to eliminate subjectivity and scale the VC investment screening process. Although several tools exist, VCs are still hesitant to adopt them, as they have difficulties comparing such novel approaches with their status quo.

Therefore, the essay describes a benchmarking exercise comparing the screening performance of VC investment professionals to that of a selected ML tool to further increase trust, accelerate the adoption of data-driven approaches and ultimately allow the traditional VC process to scale. In summary, Section 2.1 explores novel data collection approaches, Section 2.2 showcases the verification of existing datasets and Section 2.3 applies private company data to improve the VC investment process. Finally, Chapter 3 summarizes the contributions of this dissertation and suggests avenues for future research.

2 Essays

2.1 Essay 1 – How to Hit Home Runs: Portfolio Strategies and Returns in Formal and Informal Venture Capital

Abstract

We extend current research on entrepreneurial finance through the first theoretically founded and empirically tested comparison between formal and informal venture capital (VC) with respect to their investment diversification strategies and respective portfolio returns. Our novel dataset of more than 12,500 early-stage investors reveals that industry- and stage-diversification seem to drive success for both investor types. However, the underlying dependencies differ between formal and informal VCs and may even reverse the effects of diversification. In our effort to substantiate these divergent dependency effects, we validate that portfolio returns observe a power-law distribution for both investor types.

Keywords:	Venture capital, business angels, returns, diversification
Authors:	Andre Retterath, Stelios Kavadias
First Author:	Andre Retterath
Current Status:	Submitted to the Journal of Business Venturing; presented at 3rd Entrepreneurial Finance Conference (Milan, 2019); working paper available at *https://ssrn.com/abstract=3527412*

2.1.1 Introduction

Angel investors such as Peter Thiel, Mark Cuban, Dave McClure, or Reid Hoffman are well-known for their early and highly successful investments in companies like Facebook, Uber, Lyft, Airbnb and many more. Together with other, less publicly known individuals, as well as families and friends of the entrepreneurs, such angel investors form a group of informal venture capitalists (*iVCs*). Although iVC represents between 60-90% of the total capital provided to startups (Mason and Harrison, 2017; Mason, 2006; Wilson and Silva, 2013), there is scarcity of reliable information on the relationship between their investment portfolio strategies, and their respective returns. This marks a significant asymmetry between the economic importance and academic visibility of iVCs. Instead, analyses have been extensively performed for institutional players like Sequoia Capital, New Enterprise Associates, or Kleiner Perkins whom we classify as formal venture capitalists (*fVCs*).

In reality, both fVCs and iVCs face high-risk, high-return investment contexts, where they are equally incentivized to manage risk and optimize returns. Therefore, it seems straightforward to expect that research findings on the relationship between portfolio investment strategies and returns translate from fVC to iVC contexts. Yet, despite the contextual similarities, fVCs and iVCs differ in a variety of dimensions: their formal structures, governance, resources, depth of experience, and risk tolerance, among others. As such, their structural dissimilarities can provide the basis for potential differences in their investment strategies and portfolio returns.

The objective of this paper is to explore to what extent existing research findings on the relationship between fVCs' portfolio diversification or specialization strategies and their returns translate to iVCs. The academic literature on the question of investment portfolio diversification versus specialization is well established for fVCs (Bygrave, 1988; Cressy et al., 2014; Cressy, Munari, and Malipiero, 2007; Gompers et al., 2009; Gorman and Sahlman, 1989; Humphery-Jenner, 2013; Jackson III, Bates, and Bradford, 2012; Knill, 2009; Matusik and

Fitza, 2012; Norton and Tenenbaum, 1993; Sahlman, 1990). Unfortunately, scholars have been unable to tackle the equivalent strategic question for iVCs due to the lack of suitable data.

We overcome the data availability limitation and assemble an original dataset through a novel bottom-up approach of collecting, processing and verifying data for both fVC and iVC. We validate the dataset's robustness through the replication of existing results for fVC strategies before we address iVC strategies, and then we conduct a detailed comparison between both investor groups. Hereby, we provide the first theoretically founded, and empirically tested comparison between fVCs and iVCs with respect to their diversification strategies and portfolio returns. Although we are unable to empirically show whether iVCs intentionally pursue a diversification strategy, or have their portfolios passively emerge[1], our analysis provides a strategic framework that can help iVCs to actively improve their performance in the future.

In our effort to analyze these differences we explore three distinct, but closely associated, hypotheses. First, we hypothesize that, in line with the financial theory, fVCs and iVCs similarly benefit from diversification in the industry and stage of the ventures they invest in. At the same time, however, we expect that due to several conditions, iVCs naturally benefit from specialization which mitigates the effect of diversification, i.e., the positive effect of stage and industry diversification is stronger for fVCs than for iVCs (*Diversification Hypotheses*). Second, we consider the dependence of the relationship between diversification strategies and portfolio returns on previous investment experience, as well as on stage- and industry-specific risks. We conjecture that the direction and the magnitude of these effects are different for both investor types (*Dependency Hypotheses*). Third, we hypothesize that while fVCs strive to maximize their upside and thus generate power-law distributed returns, iVC returns should be more normally distributed due to their smaller portfolio size, and their incentives to minimize their downside risk (*Return Hypothesis*).

[1] Anecdotal evidence from informal conversations with several iVCs in the Cambridge (UK) start-up ecosystem indicates that both approaches can co-exist amongst different investors.

The remainder of this paper is structured as follows: Section 2.1.2 reviews the existing literature and develops the theoretical arguments about the similarities and differences between fVCs and iVCs. Subsequently, section 2.1.3 details the data collection process, the sample statistics, the research design, and the econometric methods. Section 2.1.4 presents our results and robustness tests, whereas, in section 2.1.5 we discuss our findings, contributions, implications but also limitations of our analysis.

2.1.2 Theory and hypotheses

Start-ups[2] are new entrepreneurial ventures which aim to meet specific market needs by providing an economically viable product, service, process, or platform. Most such ventures fail at very early stages of their existence. Therefore, they are investments that carry significant amounts of risk for all stakeholders involved (Ruhnka and Young, 1991). Established financial institutions have strategically kept away from such types of investments, not developing any expertise to evaluate such nascent businesses (Hellmann, Lindsey, and Puri, 2007). Thus, a growing number of alternative financing institutions, broadly summarized under the tagline "venture capitalists" (VCs) have developed the capability to systematically assess risks and opportunities faced by such ventures, and to (financially) support them. Without these VCs, new ventures are rarely able to secure the necessary resources and, thus, end up unable to pursue their vision (Gompers and Lerner, 2004; Gorman and Sahlman, 1989; Kortum and Lerner, 2001).

2.1.2.1 Formal and informal venture capital

VCs can be clustered into two distinct subgroups: *formal* VCs (*fVCs*) which are legally registered entities, managed by a group of general partners, and *informal* VCs (*iVCs*), which comprise individuals such as business angels or family and friends of entrepreneurs (Freear et al., 1994). In line with a variety of industry reports and previous research from Mason and

[2] The terms "start-up," "business," "venture," "investee" and "company" will be used interchangeably.

Harrison (2017), Mason (2006), as well as Wilson and Silva (2013), the importance of iVC steadily increases and accounts for 60-90%[3] of the invested early-stage capital. Surprisingly, however, less than a quarter[4] of the published research papers on VCs focus on iVCs, indicating a strong asymmetry between the economic importance and the academic attention to iVC activities.

Despite facing similar high-risk, high-return investment contexts, fVCs and iVCs differ in a variety of aspects. Most obviously, fVCs are registered investment entities, whereas iVCs are private and self-certified individuals. Due to their structure, fVCs have an obligation to report their performance to the external parties that provide the capital, i.e., their limited partners (*LPs*), whereas iVCs have no formal requirement to report to anyone because they invest their private capital. Moreover, the total amount of capital available for investments typically differs. fVCs tend to have significantly larger funds which allow for a larger average allocation per investment, more capital for follow-on financing rounds, and a higher number of concurrent investments compared to iVCs. Additionally, fVCs engage more experts, i.e., multiple investment managers versus one individual iVC. As such, fVCs are not only able to handle more investments but to gain deeper experience per investment manager across a variety of industries, regions, and technologies. Instead, iVCs are limited to fewer domains of expertise because of their resource constraints. For example, successful entrepreneurs who become iVCs, often focus on areas where they have previously gained operational experience. Another major difference is the investment objective. While iVCs may follow a combination of financial, personal and idealistic objectives, fVCs are predominantly financially driven. Due to the VC fund mechanisms, fVCs have strong external pressure to deploy their capital and achieve superior performance as they would otherwise not be able to raise subsequent funds or surpass

[3] Exact percentage values strongly depend on the country and measurement method.
[4] This stems from a rough approximation based on ca. 450,000 papers with keywords “informal venture capital,” “business angels” and “angel investors,” versus ca. 1,900,000 papers with keywords “formal venture capital,” “venture capital firms” and “venture capital funds” on Google Scholar. This ratio is representative for other literature databases.

the hurdle rate and receive their carried bonus. Unallocated capital, so-called dry powder, becomes an issue for fVCs towards the end of a fund duration, whereas for iVCs there is no pressure to deploy a specific amount of capital at all. Certainly, these dissimilarities offer a basis for anticipating differences in investment strategies.

In summary, both investor types face similar external contexts but are likely to behave differently based on their internal conditions. To understand how VCs manage and control their risks, we need to understand the risk construct in more detail.

2.1.2.2 Early-stage investor risk and mitigation strategies

Investment risk can be dissected into two components, the systematic, uncontrollable, market or economy-specific risk, and the unsystematic, controllable, company, or industry, asset-specific risk (Norton and Tenenbaum, 1993). Financial markets reward uncontrollable systematic risk with significantly higher returns, whereas controllable unsystematic or idiosyncratic risk is not rewarded (Ross, 1976; Sharpe, 1964). Filling a market void, VCs are incentivized to accept systematic risk and to manage unsystematic risk.

Micro and macro level risk mitigation strategies. Unsystematic, or idiosyncratic risk can be managed through two different set of approaches: a *deal-specific micro risk* approach, and a *portfolio-specific macro risk* one. On the micro-level, several survey-based papers have studied deal-specific risk mitigation practices such as special deal structures, liquidation preferences, or deal evaluation pipelines where only 1-3% of the targets receive financing (MacMillan, Kulow, and Khoylian, 1989; Sahlman, 1990; Jeffry A Timmons and Bygrave, 1986). These general practices are similarly available for both fVCs and iVCs.

On the macro level, both the strategic management literature and the financial theory have advocated portfolio-level risk mitigation strategies through the “right” balance between portfolio diversification and specialization. Early results from financial theory argue that investors in general (Markowitz, 1952) and fVCs specifically (Sharpe, 1963) can minimize the

effects of unsystematic risk through a diversified collection of assets. In contrast, management theory argues for specialization due to information sharing benefits (Norton and Tenenbaum, 1993), organizational improvements in investment capability (Gompers, Kovner, and Lerner, 2009), and a learning curve associated with industry and/or technology knowledge (Sahlman, 1990). Previously, scholars have only focused on fVC portfolio strategies and neglected iVCs due to the scarcity of robust data.

Although the relationship between portfolio strategies and investor performance is expected to vary based on the fundamental differences between iVCs and fVCs, their conceivable positions regarding macro-level risk mitigation are essentially driven by the same dimensions, i.e., diversification and specialization.

Diversification versus specialization strategies. We dissect diversification (and/or specialization) into three key dimensions: geography, industry and venture stage. Given the geographic focus of our dataset on the Cambridge (UK) ecosystem, the remainder of this section details the latter two dimensions.

Industry. VCs can choose whether to diversify across widely different industries or to specialize and focus on specific ones. The decision is multidimensional and admittedly complex, as there are high- and low-tech verticals with large and small market potentials, high or low growth expectations, and very different risk-return profiles. An extensive stream of literature highlights the costs and benefits of industry diversification for fVCs (Bygrave, 1988; Cressy et al., 2014; Cressy, Munari, and Malipiero, 2007; Gompers et al., 2009; Gorman and Sahlman, 1989; Humphery-Jenner, 2013; Jackson III, Bates, and Bradford, 2012; Knill, 2009; Matusik and Fitza, 2012; Norton and Tenenbaum, 1993; Sahlman, 1990). Still, there is divergence on whether diversification is more successful than specialization. Knill (2009), for instance, finds that diversification across industries enables fVCs to raise more capital in subsequent funds

and, consequently, deduces that it leads to better performance. Humphery-Jenner (2013) offers additional evidence for a positive relationship between industry diversification and performance, which is likely due to learning across investments and knowledge sharing. Organizational learning theory argues that contrary to specialized investors whose survival depends on exceptional success in their area of expertise, diversified VCs have less competitive pressure in one specific industry and thus better chances of survival. Additionally, diversified VCs are able to transfer learnings across similar industries which can provide a competitive advantage (Barnett, Greve, and Park, 1994; Haunschild and Sullivan, 2002; Ingram and Baum, 1997; Matusik and Fitza, 2012). On the other hand, Bygrave (1988), Norton and Tenenbaum (1993), Sahlman (1990), Timmons and Bygrave (1986), as well as Gompers et al. (2009), establish benefits from specialization, such as learning curves due to technical experience, organizational improvements or information sharing, and provide evidence for a positive relationship to performance. Specialized VCs establish an industry-specific reputation and secure proprietary deal flow.

Despite the ongoing academic debate, recent empirical literature supports the benefits of industry diversification in the case of fVCs (Buchner et al., 2017). However, assuming that the great majority of iVCs have previous operational experience and beneficial knowledge within a specific industry, the aforementioned arguments of specialization (i.e., industry-specific reputation and proprietary deal flow access) seem to counterbalance the diversification arguments. Their unique experience helps iVCs to develop better judgment, to add more value as an advisor or board member of the company, and to establish a "quality seal," which is a signaling effect towards future investors and potential customers. Furthermore, iVCs are naturally limited to fewer industries due to the aforementioned capital constraints, which theoretically mitigates the advantages of diversification. In summary, we assume a positive effect of

diversification that is stronger for fVCs than for iVCs. Thus, we pose the following hypothesis:

H1*: The impact of industry diversification on performance is positive for both investor groups, however, it is stronger for fVCs than for iVCs, i.e., fVCs benefit more from industry diversification than iVCs.*

Venture Stage. It is equally important for VCs to determine whether they should diversify across the different stages of a venture's evolution, or to specialize and focus on few. Plummer and Walker (1987), as well as Ruhnka and Young (1991), provide evidence that with increasing maturity and number of financing rounds per company, the venture-specific risk decreases. In line with findings from Norton and Tenenbaum (1993), Buchner et al. (2017) show that stage diversification leads to higher returns for fVCs. By counterbalancing their portfolios and investing in later less risky stages, fVCs limit their downside, while simultaneously engaging in earlier and riskier stages which enhances their expected returns. The authors find that stage-specific experience due to stage specialization does not affect performance. Moreover, anecdotal evidence allows us to assume that spreading initial investments across venture stages prevents a concentration of exit timings, i.e., successful portfolio companies are in different stages and likely exit at different times. Therefore, spreading across stages reduces the risk that the portfolio performance faces any temporal negative (public) market sentiment. While there seems to be no logical argument against the value of stage diversification for fVCs, iVCs are mostly confined to earlier stages due to their lower average capital availability. Thus, we expect the benefits of stage diversification to be weaker for iVCs. Therefore, we hypothesize the following:

H2*: The impact of stage diversification on performance is positive for both investor groups, however, it is stronger for fVCs than for iVCs, i.e., fVCs benefit more from stage diversification than iVCs.*

We summarize *H1* and *H2* as the ***Diversification Hypotheses.***

Portfolio strategy dependencies. Buchner et al. (2017) were among the first to provide empirical evidence on the dependency of the relationship between diversification and performance on other factors such as the VC's risk exposure, or previous investment experience. We consider three such specific moderating factors: the portfolio-related factors of stage- and industry-related risks, and the investor-related factor of investment experience in terms of the number of investments (Buchner et al., 2017).

Stage-related risk. A venture's defaulting risk decreases with a later development stage and higher maturity (Plummer and Walker, 1987; Ruhnka and Young, 1991). In line with Cumming (2006) and Buchner et al. (2017), we assume that stage-specific risk exposure has a significant impact on the relationship between diversification and performance. Building on Norton and Tenenbaum (1993), we expect diversification to be more beneficial in portfolios focused on earlier/riskier stages. This is aligned with the fundamental theoretical value of diversification as a successful investment strategy in high-risk contexts (Markowitz, 1952). There seems to be no theoretically founded argument for assuming a difference with respect to this effect for fVC and iVC.

H3a*: Target stage per investor has a similar negative impact on the relationship between industry diversification and investor performance for both fVCs and iVCs, i.e., the later (earlier) the stage, the less (more) successful is industry diversification.*

H3b*: Target stage per investor has a similar negative impact on the relationship between stage diversification and investor performance for both fVCs and iVCs, i.e., the later (earlier) the stage, the less (more) successful is stage diversification.*

Industry-related risk. We follow recent findings whereby industry-specific risk exhibits a negative impact on the relationship between diversification and performance for fVCs (Buchner et al., 2017). Cumming (2006) also provides supporting evidence

that fVCs tend to focus and specialize in fewer industries when their investments become more resource-intensive and complex in industries such as life sciences (biotechnology and medical) or software. In line with our previous argumentation that iVCs are generally better off by specializing in fewer industries, we presume that the specialization effect is even stronger in riskier industries. With respect to stage diversification, there exists to our knowledge no evidence or theoretical insight against the general understanding that diversification is more successful in riskier scenarios. Thus, we assume:

H4a*: Target industry-risk has a similar negative impact on the relationship between industry diversification and investor performance for fVCs and iVCs, i.e., the higher the industry risk, the less (more) successful is industry diversification (specialization).*

H4b: *Target industry-risk has a similar positive impact on the relationship between stage diversification and performance for fVCs and iVCs, i.e., the higher the industry risk, the more (less) successful is stage diversification (specialization).*

Investment experience. In line with the benefits of specialization, Bernile, Cumming, and Lyandres (2007), Dimov and De Clercq (2006), Jackson III et al. (2012), as well as Kanniainen and Keuschnigg (2003), argue that managing and assisting investees requires significant resources from the investor. Motivated by limited attention theories, Cumming (2006), Cumming and Dai (2011), Gifford (1997), as well as Jääskeläinen, Maula, and Seppä (2006), find that a smaller number of investments leads to a better portfolio performance as the VCs allocate more time per company. Similarly, Gompers and Lerner (1999, 2001), Kanniainen and Keuschnigg (2003, 2004), as well as Hsu (2004), argue that there is a trade-off between the quality of VC advice and the number of investees per investor. Advice and value-add tend to dilute with an increasing number of investees. To overcome this issue, we posit that fVCs retain the size of their portfolios through specialization in fewer industries, where information sharing and learning

across investments tend to be stronger (Humphery-Jenner, 2013). Eventually, more investment experience and a higher number of investments allow fVCs to establish a track record which helps them to attract more deal flow, to access better deals and, ultimately, to achieve the desired portfolio structure and superior returns (Buchner et al., 2017). We posit that these effects are stronger within industries than across. Thus, with an increasing number of investments and track record, fVCs should be specializing in fewer industries. We expect investment experience to amplify the positive relationship between industry specialization and portfolio performance for iVCs. With respect to stage diversification, we find no evidence or theoretical argument for/against the moderating effect of investment experience. Thus, we hypothesize that:

H5a*: Investor experience has a negative impact on the relationship between industry diversification and investor performance for fVCs and iVCs, i.e., the more investment experience, the less (more) successful is diversification (specialization).*

H5b*: Investor experience has no impact on the relationship between stage diversification and investor performance for fVCs and iVCs.*

We summarize hypotheses *H3* to *H5* as the ***Dependency Hypotheses***. After we have elaborated on the risk mitigation strategies, we discuss the investor performance next.

2.1.2.3 Investor/portfolio performance and returns

Investor success can be assessed along multiple dimensions. In this paper we focus on the investor financial performance[5]. As VCs invest in multiple assets in parallel, this portfolio-level metric is an aggregate of the underlying investment performances. With respect to a suitable performance metric, the literature is replete with discussions on particular advantages and disadvantages but has not yet identified the "one-measure-fits-it-all," neither on the portfolio-level nor on the investment-level (Brush and Vanderwerf, 1992; Hochberg,

[5] More formally, as per Markowitz (1952), the investors' objective is to "*maximize discounted expected, or anticipated, returns."*

Ljungqvist, and Lu, 2007; J. E. Lange, Bygrave, Nishimoto, Roedel, and Stock, 2001; Sandberg and Hofer, 1987). To summarize main issues from previous literature, the proposed measures of investor and/or investment performance are either not suitable due to their binary and mostly retrospective nature, e.g., positive exits such as IPOs and trade sales (Bottazzi et al., 2007; Gompers, Kovner, and Lerner, 2009; Hege, Palomino, Schwienbacher, et al., 2003), or not available due to the private status of both the investor and the investee, e.g., financial metrics such as a fund‘s IRR (Aghion and Bolton, 1992; Baer and Frese, 2003; Brush and Vanderwerf, 1992; Hillman and Keim, 2001; Jennings and Beaver, 1997; Sandberg and Hofer, 1987). In this paper, we focus on the relative comparison between fVCs and iVCs, rather than an absolute analysis of investment returns. Thus, an exact approximation of the absolute investor performance is subordinate to the data collection process, and less crucial than in other studies.

With respect to the return profile of fVC portfolios, the extant literature concurs that the distribution of fVC returns is highly skewed, as these are non-normally distributed and follow a *power-law*. This implies that a small number of so-called "home run investments" accounts for the majority of portfolio returns, representing the fact that fVCs seek to maximize their upside (exit potential) rather than protecting the downside (risk of bankruptcy) for every single investment (Gompers and Lerner, 1999; Kaplan et al., 2005; Korteweg and Sorensen, 2010). However, due to their limited capital availability and a smaller number of investments, it is unclear whether, and to what extent, the same logic applies to iVCs. The relatively smaller number of investments makes it more difficult to strive for home runs. Moreover, iVCs' incentives make them seek a minimized downside loss at the expense of moderate upside potential. In Mason and Harrison (2002)'s words, "*business angels therefore concentrate on avoiding bad investments rather than seeking winners and aim to make a return on every investment.*" Consequently, we assume:

H6: *fVC returns follow a power-law distribution, whereas iVC returns are more normally distributed.*

We refer to *H6* as the ***Return Hypothesis***.

2.1.3 Data and empirical methods

A major constraint in empirical research on entrepreneurial finance has consistently been the limited data availability. Across all the pre-IPO growth stages of a venture, both investors[6] and investees retain a private status. Therefore, they have no obligation to publicize internal information (Freear et al., 1994; Kaplan and Lerner, 2016; Mason and Harrison, 2002). To our knowledge, this study is the first to overcome this hurdle and provide a large-scale dataset for both fVC and iVC. Hereafter, we describe in detail our data collection process. Subsequently, we elaborate on the sample formation procedure, and we offer descriptive statistics of our final sample. Finally, we describe our research design and how it compares with previous literature.

2.1.3.1 Data collection procedure

The most common way of collecting VC data is through one of the many external databases such as Crunchbase, VentureXpert, Angellist, Tracxn, CB Insights, or Pitchbook. These aggregators collect, structure and analyze data through various streams of information such as well-connected in-house research teams, external data providers, crowdsourcing data from entrepreneurial communities, web crawlers or machine learning models. In practice, though, these databases appear to be partially correct, often incomplete, and mostly limited to fVC (Kaplan and Lerner, 2016; Kaplan et al., 2002). They rarely include information on iVCs which makes them unsuitable for our study. Similarly, survey data collection approaches such as top-down (VC portfolio data from their LPs), bottom-up (VC portfolio reverse-engineered from their respective portfolio companies), or VC directly are unsuitable for collecting a large-scale dataset which allows us to compare fVC and iVC to rigorously test our hypotheses.

[6] Publicly listed fVCs exist but are negligible, i.e., less than 5% based on 241 directly listed VC firms versus 5,049 private VC firms worldwide in 2018 identified through Crunchbase. For comparison, Pitchbook provides a ratio of 4.3%. iVCs are private per definition.

We circumvent the respective challenges and assemble a novel dataset by combining three simple concepts. First, we achieve scalability through the collection of all shareholder information for a specific window of time, and a contained geography through an official government company register. Second, we cover the investments from the iVCs almost in their entirety, by complying with the literature assumption that the great majority of iVCs invest close to their home, and within their familiar ecosystems[7] (Freear et al., 1994). Third, we follow a bottom-up collection procedure, automatically clean the resulting dataset and reverse-engineer the VC portfolios. While we can be certain that under these conditions, we capture the full portfolios of the iVCs, we likely miss some of the fVCs' investments, due to their activities outside of our selected geography. We discuss this potential shortcoming in Section 2.1.5.

The United Kingdom's (UK) government registrar of companies, the "Companies House" (*CH*), incorporates and dissolves limited companies within the UK. They examine and store basic company information such as shareholder structures or capital modifications and make this information publicly available. CH requires the companies to regularly provide (oftentimes handwritten) forms which are then scanned and published on their website. A private publisher, "Bureau Van Dijk" (*BVD*), employs around 900 professionals to scrape, structure and enrich this data with additional information. For the UK and Ireland, their database "Fame" provides more than 20 years of detailed information on over 11 million active and inactive companies.

Based on the above-described data availability, we sought a geographically definable, but highly active ecosystem, specifically with respect to iVCs. The entrepreneurial cluster of Cambridge, UK, is one of the densest and most active ecosystems in Europe and, thus, it provides a unique opportunity for our bottom-up data collection. We consider a ten-year period between January 1st, 2007 and December 31st, 2016 as our data collection window. This

[7] In this particular case the chosen ecosystem is one of the major ones within the UK (i.e., Cambridge), and as such we posit an even stronger attraction of the iVCs on their familiar ecosystem.

timeframe is suitable as it captures at least one full investment cycle from VCs[8]. The combination of CH and BVD provides data on 18,245 investors who supported 17,840 companies (8,921 active, 6,266 dissolved, 1,233 being inactive but not dissolved, and 1,420 without status) with 20,641 investments across all 355 industries documented in the database.

We source the company data from the BVD database as they have corrected the majority of spelling errors in CH, and they provide all relevant information in a structured digital form. To verify the data and reach "truth-level" quality, we follow a common supervised machine learning approach known as rule-based data mining (Han et al., 2011). We select a sub-sample from the respective BVD dataset for which we have proprietary access to the *true* company information[9] such as shareholder structure, date of incorporation and date of financing investments. We then manually clean the selected sub-sample by matching the BVD data with the *true* data and replacing it wherever differing. Every manual edit leads the supervised algorithm to create a new rule. Once completed, we automatically execute the resulting ruleset to clean the full BVD dataset in line with our manual changes. It results in our final sample of *12,588 investors who supported 3,328 Cambridge-based companies through 14,575 investments*.

The final dataset represents with very high accuracy the entire active shareholder landscape of the Cambridge ecosystem for the respective time window. It includes families, friends, professional angels, accelerators, university endowments, seed funds and growth funds, among others. The active investor base, i.e., those with three or more investments, is split into 85.78% iVC and 14.22% fVC, with 53.17% and 46.83% of the number of investments respectively. This supports Mason and Harrison's (2017), Mason's (2006), and Wilson and Silva's (2013) findings that iVCs represent between 60-90% of the total capital provided to

[8] We assume that fVCs follow investment cycles of seven to maximum ten years and, thus, a suitable timeframe should capture at least one full investment cycle, i.e., ten years. The timeframe is less relevant to iVCs as they continuously invest from their private savings, similar to an evergreen fund structure.

[9] The access was provided through a variety of approaches, i.e., internal databases, reaching physically out to the founders/investors, and desk research.

startups. We analyze the full sample and various investment-activity-based subsamples to test for robustness. A comparison to other studies is not feasible as, to our knowledge, this is the first dataset of its kind.

2.1.3.2 Variables and summary statistics

Dependent variable. By virtue of our bottom-up data collection approach, we cannot use a straightforward dependent variable like the internal rate of return (IRR) to measure investor performance at the portfolio level. However, for the purpose of our study, a reliable measure that captures performance comparisons across investors suffices. To build an investor performance measure at a portfolio level, we first focus on the individual performances of the underlying investments of the portfolio.

We ensure the validity of our measure by assuming that the financial performance of individual investments is positively related to the change in the valuation of the underlying ventures, for example, the larger the valuation gains of a venture, the better the respective investment performance. In fact, Paul's (2016) analysis of over 5,000 capitalization tables, finds an exponential relationship ($y = 7.3006\ e^{0.5294x}$) between the number of financing rounds (x) and the mean venture valuation (y), and a linear relationship ($z = 14.849x - 14.686$) between the number of financing rounds (x) and the median venture valuation (z). His respective R-squared values of 0.9619 and 0.9838 indicate a great fit. The fact that the mean valuations exceed the median ones indicates significant positive outliers, and a skewed distribution[10]. Thus, a higher number of financing rounds makes a company (exponentially) more valuable. This observation echoes past literature wherein scholars define venture success as the "*ability to attract an additional round of financing*" (Alexy et al., 2012; Hochberg et al., 2007; Ter Wal et al., 2016).

[10] Multiple reports from dealroom.co, Pitchbook, and CB-Insights on the annual development of venture valuations across financing rounds confirm both relationships with similar coefficients.

We test this conjecture by identifying all companies with an IPO as their exit outcome within our dataset, for which we extract their respective total number of financing rounds (*Tot*). We find that all IPO companies have their *Tot* in the top decile of all companies' *Tot,* i.e., a positive relationship between *Tot* and venture valuation. Since the total number of financing rounds is positively[11] related to a venture's valuation, which is also positively related to an investment's performance, we credibly posit that the total number of financing rounds positively relates to an investment's performance. These relationships hold independent of the investor type.

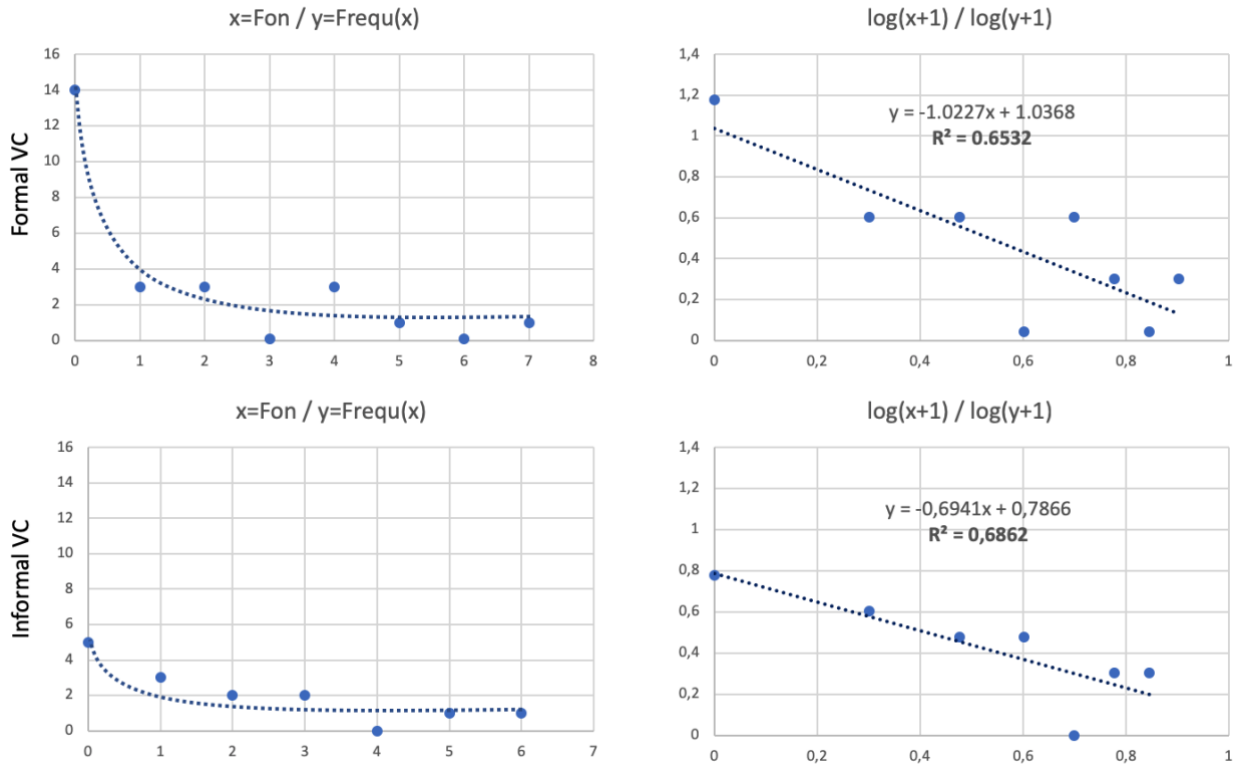
Moreover, we note that most investors would not invest at the day of a venture's incorporation but only appreciate its valuation gains after their initial investment. To account for this reality, we adopt a tweaked measure for an investment's performance: the number of financing rounds a venture receives *after* the initial investment of an investor, i.e., the *number of follow-on financing rounds (Fon)*. In summary, the more financing rounds happen after an investor's initial investment, the better the performance of the respective investment.

While we have established quantitative evidence on the relationship between the follow-on number of financing rounds and the median/mean venture valuation, we need to further understand how these individual investments drive the compound portfolio returns for both investor types in order to produce a suitable portfolio level measure. For that reason, we seek to understand the structure of the return distributions, as these reveal which investments shape the total return. Figure 2.1.1 illustrates randomly selected portfolio distributions for one fVC (top left) and one iVC (bottom left). They are representative for the active investor sample.

[11] We conjecture a linear relationship for median valuations and an exponential relationship for mean venture valuations. This differentiation needs to be explicitly considered when interpreting the OLS and Poisson coefficients.

Figure 2.1.1: Power-law distribution of portfolio returns for formal VC and informal VC

This figure depicts the power-law distribution of portfolio returns for formal VC in the upper two graphs and for informal VC in the lower two graphs. The x-axis represents *Fon* and the y-axis represents the frequency of the respective *Fon* within the investor's portfolio. The graphs on the left exhibit a linear scale, whereas the graphs on the right exhibit a logarithmic scale.



We find that the top quartile[12] of fVCs has on average 46.32% of write-offs, i.e., investments with 0 *Fon*, whereas the top quartile of iVCs has on average 35.48% of write-offs[13]. For the depicted examples in Figure 2.1.1, the fVC has 14/25 (56.00%) of the investments with 0 *Fon* whereas the iVC has 5/14 (35.71%) of the investments with 0 *Fon*. Additionally, we find that *the maximum number of follow-on financing rounds across the VC's portfolio* (*MaxFon*) within the top quartile of fVCs is on average 13.86% higher than for the top quartile of iVCs. The respective ratio of *MaxFon* fVC versus *MaxFon* iVC in the illustrated example is 7/6, i.e., the fVC is 16.67% more successful based on *MaxFon*. These observations

[12] We select the top quartile with respect to *Fon* as we are interested in a sufficient sample size of the best-performing investors. The magnitude of the respective effects is even stronger in the top decile.
[13] We exclude positive exits so that 0 *Fon* represents write-offs, i.e., unsuccessful companies, often termed "walking dead."

offer structural insights into the return distributions of different types of investors and essentially test our ***Return Hypothesis H6***. They support our assumption that iVCs tend to minimize their downside risk rather than maximizing their upside potential, and that fVCs maximize their upside potential rather than protecting their downside risk. Yet, they also show that independently of the investor type, the distribution of returns is right-skewed and deviates significantly from a normal distribution.[14]

We test our ***Return Hypothesis*** formally for the described sample by calculating the individual p-values based on the respective Pearson correlation coefficients of the log-log transformed *Fon* distributions. These range from 0.7723 to 0.8943, and the respective sample sizes, i.e., the number of investees per investor ranges from 5 to 23. Hereby, we end up with p-values ranging from 0.000001 to 0.127745 for both investor types. 97.3% and 95.2% of the analyzed log-log transformed *Fon* distributions for fVCs and iVCs respectively return a p-value <0.1, which rejects the null hypothesis that returns are not power-law distributed for both investor groups on a 10% level. Moreover, 53.2% of fVCs and 44.3% of iVCs return a p-value <0.01 which rejects the null hypothesis for approximately half of both investor groups on a 1% level. Consequently, we reject H6.

In line with the previous literature, and given that the performance of power-law distributed portfolios depends on their most successful outlier, i.e., their home run investment (Kaplan et al., 2005; Sahlman, 1990), we select the *maximum number of Fon* (*MaxFon*) as our dependent variable. We note again that our approximation is sufficient for a quantified comparison across investors.

Independent variables. We measure two dimensions of diversification for the investor's investment strategies: industry-specific and stage-specific. We build upon the previous literature (Buchner et al., 2017; Cressy et al., 2014; Dimov and De Clercq, 2006; P. Gompers

[14] Figure 1 also shows the Log-Log transformed graphs for fVC (top right) and iVC (bottom right) in a linear shape. R-squared values for a sample of 50 fVCs and 50 iVCs with a sufficient number of investments range from 60.52-80.21% and provide strong evidence for a power-law distribution for both investor types.

et al., 2009; Jääskeläinen et al., 2006; Yang, Narayanan, and De Carolis, 2014) to define the following measures: macro industry diversification (MacIndDivk) as {1 – Herfindahl15 index of the different industries represented in the portfolio of investor k}, and stage diversification (StaDivk) as {1 – Herfindahl index of the different initial investment stages represented in the portfolio of investor k}. We classify the industries based on the first digit of the SIC code which leads to 21 macro industries, that ensure a sufficient sample size.

Moderator variables. We include each investor's investment experience as a moderator variable and measure it as the *total number of investments per investor* (Inv_k). This is a count variable, which in our sample ranges from 1 to 34. Besides this investor specific factor, we also assume that the relationship between the independent variables and *MaxFon* might be impacted by several asset-related external factors. Therefore, we consider the average industry riskiness for an investor *k* ($AvgIndRis_k$) and the average stage or financing round for an investor *k* ($AvgSta_k$) as additional moderator variables. The latter is straightforward as the risk typically decreases with an increasing number of financing rounds. It is a continuous variable and it is calculated by the sum of the stages of the VCs initial investments divided by the number of investments per VC and ranges from 1.0 to 8.0 in our sample. With respect to $AvgIndRis_k$, however, there seem to be multiple ways of determining the industry-specific risk ranging from binary measures such as high-tech versus low-tech, or IT versus non-IT, to continuous measures based on the respective costs of experimentation. Due to the richness of our dataset and in line with the definition of our dependent variable, we decide to measure $AvgIndRis_k$ for an investor *k* as the weighted average survival rate of ventures across the respective industries:

$$AvgIndRis_k = \frac{\sum_{i=1}^{N}\left(I_{i,k}\left(1-\frac{\sum_{j=2}^{M_i}\frac{x_{i,j}}{x_{i,j-1}}}{M_{i-1}}\right)\right)}{\sum_{i=1}^{N} I_{i,k}} \quad (1)$$

[15] The Herfindahl-Hirschman Index (HHI) is a common concentration measure. It is calculated by squaring the share of portfolio companies in every industry and summing the resulting numbers.

where N represents the number of different industries for an investor k, Ii,k represents the number of investments in an industry i from an investor k, Mi represents the maximum number of financing rounds in an industry i and where xi,j represents the number of companies in an industry i which received a financing round in a stage j. The inner brackets represent the survival rate from one to the next stage within a specific industry. AvgIndRisk is a continuous variable that ranges from 0 to 1 with 1 representing the maximum risk. It is important to note that we remove all positive exits before calculating the survival rates as they would otherwise be considered as dropouts or non-survivors.

Control variables. In line with previous studies, we include a variety of control variables to minimize concerns of omitted variable bias and improve the specification of our model. We initially included 12 different variables but eventually removed those with significant multi-correlation and those with insignificant p-values. Therefore, we end up controlling for potential effects driven by the year of incorporation of the investee, the year of the initial investment of the investor in the investee, the year of the last investment from any investor in the investee and the number of co-investors, all for the company/investment with the *MaxFon* of a specific investor.

Summary statistics. Table 2.1.1 summarizes all variables as described above. Table 2.1.2 exhibits the descriptive statistics of our final sample along the company, investment, and investor-level. Regarding the companies, we see that the majority of the investees operates in two verticals, i.e., SIC10 (Information and communication) and SIC13 (Professional, scientific and technical activities), two high-risk sectors closely related to the research activities of the local university, i.e., University of Cambridge. Furthermore, we see an increasing number of incorporations over time as the number of new companies has almost tripled from 2007 to 2016. This is likely connected to the general economic upswing after the financial crisis and a possible result of the launch of multiple accelerators and entrepreneurship centers in the Cambridge area.

Table 2.1.1: Variable overview and description

Overview and description of the dependent variable, the two independent variables, the four moderator variables and the three control variables.

Variable Type	Data Type	Abbreviation	Description
Dependent	Ordinal	MaxFon	Maximum number of follow on investments per investor
Independent	Continuous	MacIndDiv	Herfindahl-Hirschmann Index for macro industry diversification per investor, 1=fully diversified
Independent	Continuous	StaDiv	Herfindahl-Hirschmann Index for stage diversification per investor, 1=fully diversified
Moderator	Continuous	AvgSta	Average stage of initial investment per investor
Moderator	Continuous	AvgIndRis	Average industry risk per investor, 1=maximum risk
Moderator	Dichotomous	InvTyp	Type of investor, 0=individual, 1=fund
Moderator	Ordinal	Inv	Absolute number of investments per investor as measure for portfolio size
Control	Ordinal	YeaIncIni	Years between incorporation and initial investment for company/investment with maximum number of follow on investments per investor
Control	Ordinal	YeaIniLas	Years between initial and last investment for company/investment with maximum number of follow on investments per investor
Control	Ordinal	CoIn	Number of co-investors at initial investment for company/investment with maximum number of follow on investments per investor

Table 2.1.2: Descriptive statistics

Overview of the descriptive statistics across the company, investment and investor-level. On the company-level, the table distributes the companies based on their SIC industry verticals and founding years (Incorp. Year). On the investment-level, the table distributes the investments based on the SIC industry vertical of the underlying company, the year of the initial investment of the investor in the company (YeaIni) and the initial stage of the company at the initial investment of the investor (StaIni). On the investor-level, the table distributes the investors based on their maximum number of follow-on investments (MaxFon) and based on the number of their investments (Inv).

Company-level																				
SIC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequ.	46	3	158	8	7	340	241	36	122	555	80	171	933	266	4	103	170	72	80	52
Percent	1,33	0,09	4,58	0,23	0,20	9,86	6,99	1,04	3,54	16,10	2,32	4,96	27,07	7,72	0,12	2,99	4,93	2,09	2,32	1,51
Cum	1,33	1,42	6,01	6,24	6,44	16,30	23,30	24,34	27,88	43,98	46,30	51,26	78,33	86,05	86,16	89,15	94,08	96,17	98,49	100,00
Incorp. Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016										
Frequ.	210	185	203	253	291	348	445	452	455	605										
Percent	6,09	5,37	5,89	7,34	8,44	10,10	12,91	13,11	13,20	17,55										
Cum	6,09	11,46	17,35	24,69	33,13	43,23	56,14	69,25	82,45	100,00										
Investment-level																				
formal VC																				
SIC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequ.	12	0	140	14	0	55	29	4	12	331	51	81	496	113	0	29	31	7	10	74
Percent	0,81	0,00	9,40	0,94	0,00	3,69	1,95	0,27	0,81	22,23	3,43	5,44	33,31	7,59	0,00	1,95	2,08	0,47	0,67	4,97
Cum	0,81	0,81	10,21	11,15	11,15	14,84	16,79	17,06	17,86	40,09	43,52	48,96	82,27	89,86	89,86	91,81	93,89	94,36	95,03	100,00
YeaIni	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017										
Frequ.	47	47	83	106	163	177	239	251	354	22										
Percent	3,16	3,16	5,57	7,12	10,95	11,89	16,05	16,86	23,77	1,48										
Cum	3,16	6,31	11,89	19,01	29,95	41,84	57,89	74,75	98,52	100,00										
StaIni	1	2	3	4	5	6	7	8	9	10										
Frequ.	7367	3423	1018	583	312	199	110	54	8	2										
Percent	56,34	26,18	7,79	4,46	2,39	1,52	0,84	0,41	0,06	0,02										
Cum	56,34	82,52	90,30	94,76	97,15	98,67	99,51	99,92	99,98	100,00										
informal VC																				
SIC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequ.	110	5	930	59	21	911	681	123	315	2378	268	941	3114	777	19	284	504	201	193	1242
Percent	0,84	0,04	7,11	0,45	0,16	6,97	5,21	0,94	2,41	18,19	2,05	7,20	23,81	5,94	0,15	2,17	3,85	1,54	1,48	9,50
Cum	0,84	0,88	7,99	8,44	8,60	15,57	20,78	21,72	24,13	42,31	44,36	51,56	75,37	81,32	81,46	83,63	87,49	89,03	90,50	100,00
YeaIni	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017										
Frequ.	440	535	825	1173	1760	1600	1837	1974	2812	120										
Percent	3,36	4,09	6,31	8,97	13,46	12,24	14,05	15,10	21,51	0,92										
Cum	3,36	7,46	13,77	22,74	36,20	48,43	62,48	77,58	99,08	100,00										
StaIni	1	2	3	4	5	6	7	8	9	10										
Frequ.	511	499	206	97	75	64	17	10	9	1										
Percent	34,32	33,51	13,83	6,51	5,04	4,30	1,14	0,67	0,60	0,07										
Cum	34,32	67,83	81,67	88,18	93,22	97,52	98,66	99,33	99,93	100.00										
Investor-level																				
formal VC																				
MaxFon	0	1	2	3	4	5	6	7	8											
Frequ.	421	245	132	67	65	30	24	16	2											
Percent	42.02	24.45	13.17	6.67	6.49	2.99	2.40	1.60	0.20											
Cum.	42.02	66.47	79.64	86.33	92.81	95.81	98.20	99.80	100.00											
Inv	1	2	3	4	5	6	7	8	9	10	12	13	18	23	25	34				
Frequ.	826	95	31	14	14	5	2	2	4	2	1	2	1	1	1	1				
Percent	82.44	9.48	3.09	1.40	1.40	0.50	0.20	0.20	0.40	0.20	0.10	0.20	0.10	0.10	0.10	0.10				
Cum.	82.44	91.92	95.01	96.41	97.80	98.30	98.50	98.70	99.10	99.30	99.40	99.60	99.70	99.80	99.90	100.00				
informal VC																				
MaxFon	0	1	2	3	4	5	6	7	8											
Frequ.	6499	2412	1004	624	524	280	146	69	18											
Percent	56.14	20.84	8.67	5.39	4.53	2.42	1.26	0.60	0.16											
Cum.	56.14	76.98	85.65	91.04	95.57	97.99	99.25	99.84	100.00											
Inv	1	2	3	4	5	6	7	8	9	10	11	12	13	15						
Frequ.	10507	831	161	41	12	11	1	1	4	1	3	1	1	1						
Percent	90.77	7.18	1.39	0.35	0.10	0.10	0.01	0.01	0.03	0.01	0.03	0.01	0.01	0.01						
Cum.	90.77	97.94	99.33	99.69	99.79	99.89	99.90	99.91	99.94	99.95	99.97	99.98	99.99	100.00						

From an investment viewpoint, we see that the majority of financing rounds happened in SIC10 and SIC13, in line with the number of incorporations. Surprisingly, we find that 56.34% of fVCs and only 34.32% of the iVCs invest in the first financing round of a venture. This is likely due to the early support from University-related fVC vehicles (i.e., endowment funds) and accelerators on the one hand, and a relatively sizeable number of the so-called "super angels," who are able to allocate more capital per investment and, thus, place initial investments in later rounds.

On the investor-level, we find that 42.02% of fVCs and 56.14% of iVCs have zero follow-on financing within their portfolio, which is a clear indication for the high downside risk related to early-stage investments. However, this ratio changes to an average of 46.32% for fVCs and 35.48% for iVCs once we focus on the top quartile of the active investor sample. This shift supports our previous insights that successful iVCs are capable of significantly reducing their downside risk, whereas there is a negligible difference in write-offs between successful and unsuccessful fVCs.

Table 2.1.3 illustrates the respective summary statistics and correlation matrices of our final sample. Although the magnitude of the respective correlation coefficients differs for fVC and iVC, the general relationships are similar. Most notably, we find a correlation of 0.5497 for fVC and 0.3729 for iVC between the independent variables *MacIndDiv* and *StaDiv* which indicates that investors concurrently diversify across both dimensions.

Table 2.1.3: Summary statistics of final sample

This table shows the summary statistics and correlation matrices of the final sample. Panel A and Panel B describe the formal VCs (fVC), whereas Panel C and Pabel D describe the informal VCs (iVC). Panel A and Panel C show the summary statistics including the sample size (N), the mean, median, standard deviation (SD), as well as the minimum and maximum values across all variables. Panel B and Panel D show the respective correlation matrices across all variables

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Descriptive statistics											
N		1,002	1,002	1,002	1,002	1,002	1,002	1,002	1,002	1,002	1,002
Mean		1.389	0.0297	0.0400	0.0675	0.550	2.228	1.486	1.992	1.757	10.76
Median		1	0.0000	0.0000	0.0000	0.5677	2	1	1	1	4
SD		1.734	0.122	0.137	0.179	0.103	1.405	1.972	2.182	2.132	14.87
Min		0	0	0	0	0.360	1	1	0	0	1
Max		8	0.800	0.840	0.778	0.787	9	34	9	8	75
Panel B: Correlation matrix											
MaxFon	(1)	1									
MicIndDiv	(2)	0.0621	1								
MacIndDiv	(3)	0.2785	0.0212	1							
StaDiv	(4)	0.3724	0.0211	0.5497	1						
AvgIndRis	(5)	-0.1743	-0.0194	0.0030	-0.0424	1					
AvgSta	(6)	-0.0169	-0.0031	0.0478	0.1239	-0.1496	1				
Inv	(7)	0.3041	-0.0042	0.5561	0.5219	-0.0061	0.0495	1			
YeaIncIni	(8)	-0.0631	0.0028	-0.0198	0.0005	-0.1159	0.8052	-0.0504	1		
YeaIniLas	(9)	0.8985	0.0526	0.2289	0.3060	-0.1311	-0.1101	0.2608	-0.0913	1	
CoIn	(10)	0.3038	0.1164	0.0908	0.0625	-0.2173	0.1451	0.0457	0.0635	0.1884	1

		(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Panel C: Descriptive statistics											
N		11,576	11,576	11,576	11,576	11,576	11,576	11,576	11,576	11,576	11,576
Mean		0.975	0.0293	0.0207	0.0300	0.566	1.739	1.130	1.470	1.364	9.026
Median		0	0.0000	0.0000	0.0000	0.568	1	1	1	0	3
SD		1.505	0.123	0.101	0.122	0.118	1.142	0.534	1.838	2.018	13.61
Min		0	0	0	0	0.200	1	1	0	0	1
Max		8	0.889	0.750	0.800	0.787	9	15	9	8	75
Panel D: Correlation matrix											
MaxFon	(11)	1									
MicIndDiv	(12)	-0.0044	1								
MacIndDiv	(13)	0.0677	0.0019	1							
StaDiv	(14)	0.2445	-0.0076	0.3729	1						
AvgIndRis	(15)	-0.3134	0.0054	0.0176	-0.1005	1					
AvgSta	(16)	0.1642	-0.0142	-0.0300	0.1427	-0.2850	1				
Inv	(17)	0.2068	-0.0014	0.5880	0.6649	-0.0439	0.0556	1			
YeaIncIni	(18)	0.1548	-0.0158	-0.0158	0.0730	-0.2209	0.7892	0.0103	1		
YeaIniLas	(19)	0.8684	0.0012	0.0617	0.2183	-0.2293	0.0678	0.1899	0.1044	1	
CoIn	(20)	0.4986	-0.0057	-0.0107	0.1020	-0.3497	0.3433	0.0627	0.2698	0.3341	1

2.1.4 Analysis, results and robustness tests

We begin our analysis with a replication of existing results for fVCs. This provides validity on the suitability of our novel dataset and performance measure for studying VC portfolio strategies. Subsequently, we analyze iVC strategies to record and understand differences. We run a heteroskedasticity-robust ordinary least squares (OLS) regression. However, this setup partially violates the Gauss-Markov assumptions, since the residuals are not normally distributed, there is presence of heteroscedasticity, and our dependent variable is

a count variable, distributed based on a power-law. Thus, we test the model's potentially limited efficiency through an appropriate robustness check, i.e., a Poisson estimation.

2.1.4.1 From fVC to iVC: OLS estimation on the "Diversification Hypotheses"

We closely mirror Buchner et al. (2017)'s analysis and examine the relationship between performance and Diversification as follows:

$$MaxFon_{mn} = \alpha_{mn} + \beta_{mn} Diversification_{mn} + \beta_{mn} X_{mn} + \varepsilon_{mn} \quad (2)$$

where *MaxFon* reflects the performance for fVC (*m=0*) or iVC (*m=1*), where *Diversification* represents either *MacIndDiv (n=0)* or *StaDiv (n=1)* and where *X* represents the vector of control variables as described in Table 2.1.1.

Table 2.1.4: Baseline analysis of impact of diversification on performance

This table shows OLS regression results. All variables are defined in Table 2.1.1. The dependent variable in all models is the maximum number of follow-on financing rounds (MaxFon). Models (1) and (2) represent formal VCs, whereas models (3) and (4) represent informal VCs. Models (1) and (3) exhibit the isolated relationship between industry diversification (IndDiv) and the maximum number of follow-on financing rounds (MaxFon). Models (2) and (4) exhibit the isolated relationship between stage diversification (StaDiv) and the maximum number of follow-on financing rounds (MaxFon). The values in parentheses are t-test values, based on heteroskedasticity-robust standard errors. ***, **, and * indicate significance at 1%, 5% and 10%, respectively.

VARIABLES	Ordinary Least Squares Regression			
	MaxFon fVC		MaxFon iVC	
	(1)	(2)	(3)	(4)
	IndDiv	StaDiv	IndDiv	StaDiv
MacIndDiv	0.531**		-0.0831	
	(2.009)		(-0.963)	
StaDiv		0.880***		0.462***
		(4.854)		(5.638)
Inv	0.0468***	0.0285**	0.133***	0.0551***
	(2.912)	(2.310)	(6.252)	(3.066)
YeaIncIni	0.00740	0.00481	0.00831***	0.00675***
	(1.190)	(0.759)	(3.742)	(3.030)
YeaIniLas	0.691***	0.680***	0.583***	0.580***
	(29.46)	(28.50)	(71.58)	(70.90)
CoIn	0.0160***	0.0162***	0.0256***	0.0256***
	(11.07)	(11.30)	(39.59)	(39.69)
Constant	-0.103***	-0.0920***	-0.211***	-0.133***
	(-4.256)	(-4.272)	(-9.595)	(-6.906)
Observations	1,002	1,002	11,576	11,576
Adjusted R-squared	0.832	0.836	0.805	0.806

Table 2.1.4 reports the results of Eq. (1). The coefficient of *MacIndDiv* for fVCs in model (1) is 0.531 and statistically significant at a 5% level, whereas it is -0.083 and statistically insignificant for iVCs (model (3)). Thus, we confirm existing literature on industry diversification of fVCs', but we find no relationship for iVCs. Therefore, we reject H1 for now. The coefficient of *StaDiv* for fVCs in model (2) is 0.880 and statistically significant at a 1% level, whereas it is 0.462 and statistically significant at a 1% level for iVCs (model (4)). Hereby, we confirm existing literature on stage diversification of fVCs and find a positive but half as strong relationship for iVCs. This leads us to confirm H2. In conclusion, and in line with the extent literature[16], we find that fVCs benefit from diversification and, therefore, replicate literature results with a noval dataset.

2.1.4.2 From fVC to iVC: OLS estimation on the "Dependency Hypotheses"

Echoing Buchner et al. (2017), we also explore how the effects of industry and tage diversification depend on contextual factors. More precisely, we analyze how the relationship between diversification and investor performance changes from low risk to high risk target industries, from earlier to later target stages of the initial investment and with more or less investment experience per investor. Based on Eq. (2), we interact *Diversification* with the different *Moderators*, i.e., *AvgIndRis*, *AvgSta*, and *Inv*, as follows:

$$MaxFon_{mn} = \alpha_{mn} + \beta_{mn} Diversification_{mn} * Moderators_{mn} + \beta_{mn} X_{mn} + \varepsilon_{mn} \quad (3)$$

Table 2.1.5 reports the results of Eq. (3). Our findings re-confirm H2 based on the comparison of the general effects of *StaDiv* in models (1) and (3) with the equivalent results of Table 4. However, we find that the coefficient for *MacIndDiv* for iVCs in model (3) is now 0.291 and statistically significant at a 1% level. Taken together with the respective coefficient for fVCs (0.606), these findings lead us to ultimately accept H1. Notably, the magnitude of

[16] Buchner et al. (2017), Bygrave (1988), Cressy et al. (2014), Cressy, Munari, and Malipiero (2007), Gompers et al. (2009), Gorman and Sahlman (1989), Humphery-Jenner (2013), Jackson III, Bates, and Bradford (2012), Knill (2009), Matusik and Fitza (2012), Norton and Tenenbaum (1993), and Sahlman (1990)

these effects is almost twice as large for fVCs compared to iVCs. In summary, in this more elaborate analysis, we show that both for fVCs and iVCs their industry and stage diversification investment strategies are more successful than the respective specialization counterparts. In line with our hypotheses, these effects are approximately twice as strong for fVCs than for iVCs.

Table 2.1.5: Impact of moderating effects on the relationship between diversification and performance

This table shows twofold interacted OLS regression results. All variables are defined in Table 2.1.1. The dependent variable in all models is MaxFon. Models (1) and (2) represent formal VCs, whereas models (3) and (4) represent informal VCs. Models (1) and (3) exhibit the interacted relationship between industry diversification (IndDiv) and the maximum number of follow-on financing rounds (MaxFon). Models (2) and (4) exhibit the interacted relationship between stage diversification (StaDiv) and the maximum number of follow-on financing rounds (MaxFon). The values in parentheses are t-test values, based on heteroskedasticity-robust standard errors. ***, **, and * indicate significance at 1%, 5% and 10%, respectively.

VARIABLES	Ordinary Least Squares Regression			
	MaxFon fVC		MaxFon iVC	
	(1)	(2)	(3)	(4)
	IndDiv	StaDiv	IndDiv	StaDiv
MacIndDiv	0.606**		0.291***	
	(2.233)		(3.119)	
StaDiv		0.778***		0.157*
		(3.958)		(1.728)
AvgSta	0.160***	0.140***	0.0675***	0.0547***
	(6.885)	(6.441)	(9.563)	(8.198)
AvgIndRisSur	-0.503**	-0.460**	-0.702***	-0.686***
	(-2.266)	(-2.101)	(-12.43)	(-12.21)
Inv	0.101**	0.0230	0.171***	0.0489**
	(2.427)	(0.913)	(6.563)	(2.055)
MacIndDiv x AvgIndRis	-1.511		-1.927**	
	(-0.589)		(-2.283)	
MacIndDiv x AvgSta	0.0556		0.464***	
	(0.304)		(4.479)	
MacIndDiv x Inv	-0.172**		-0.240***	
	(-2.122)		(-4.283)	
StaDiv x AvgIndRis		-0.695		-2.154***
		(-0.420)		(-3.541)
StaDiv x AvgSta		-0.105		0.112*
		(-1.242)		(1.883)
StaDiv x Inv		0.00477		0.0527
		(0.0664)		(0.949)
YeaIncIni	-0.0771***	-0.0700***	-0.0261***	-0.0229***
	(-5.299)	(-4.944)	(-7.794)	(-6.992)
YeaIniLas	0.694***	0.689***	0.579***	0.578***
	(30.07)	(29.55)	(72.25)	(71.74)
CoIn	0.0139***	0.0142***	0.0232***	0.0231***
	(9.987)	(10.19)	(36.53)	(36.26)
Constant	-0.0551	0.0233	0.105**	0.238***
	(-0.389)	(0.177)	(2.336)	(5.475)
Observations	1,002	1,002	11,576	11,576
Adjusted R-squared	0.839	0.841	0.809	0.809

The interaction effect of *MacIndDiv* and *AvgSta* for fVCs in model (1) is 0.056 and statistically insignificant, whereas it is 0.464 and statistically significant at a 1% level for iVCs (model (3)). This prompts us to reject H3a. Similarly, the interaction effect of *StaDiv* and *AvgSta* for fVCs in model (2) is -0.105 and statistically insignificant, whereas it is 0.112 and statistically significant at a 10% level for iVCs (model (4)); this makes us reject H3b. The interaction effect of *MacIndDiv* and *AvgIndRis* for fVCs in model (1) is -1.511 and statistically insignificant, whereas it is -1.927 and statistically significant at a 1% level for iVCs (model (3)), which rejects H4a based on the lack of significance on fVCs. The interaction effect of *StaDiv* and *AvgIndRis* for fVCs in model (2) is -0.695 and statistically insignificant, whereas it is -2.154 and statistically significant at a 1% level for iVCs in model (4) which rejects H4b. The interaction effect of *MacIndDiv* and *Inv* for fVCs in model (1) is -0.172 and statistically significant at a 5% level, whereas it is -1.927 and statistically significant at a 5% level for iVCs in model (3); this confirms H5a. The interaction effect of *StaDiv* and *Inv* for fVCs in model (2) is 0.005 and statistically insignificant, whereas it is 0.053 and statistically insignificant for iVCs (model (4)) which confirms H5b.

2.1.4.3 Robustness tests

We performed a variety of robustness checks on the above analyses to ensure the reliability and validity of our findings. As discussed, the OLS estimation partially violates the Gauss-Markov assumptions. Given that the dependent variable follows a non-normal distribution and is a count-type measure, a Poisson analysis would theoretically lead to more efficient estimation compared to an OLS regression. Consequently, we run the Poisson analysis, and we calculate the average partial effects to achieve comparability between the results of the OLS and the Poisson regression (Wooldridge, 2016). The coefficients including their signs and significance levels remain similar with respect to the *Diversification Hypotheses* and the *Dependency Hypotheses*. Table 2.1.6 compares the predicted values of both regressions and indicates that the adjusted R-squared values are considerably higher for the OLS than for the

Poisson regression, i.e., between 2.9 (model (3a) versus model (3b)) and 12.7 (model (1a) versus model (1b)) percentage points. More importantly, however, we see that across models the predicted values and summary statistics of the OLS estimation are almost identical to the actual data, whereas they strongly differ for the Poisson estimation, especially for fVCs. For instance, the predicted maximum value for *MaxFon* of fVCs in models (3b) and (4b) exceeds the actual value by relatively up to 70.75% which can be explained through an "over-exponentialization"[17]. To summarize, both methods lead to almost identical results in terms of coefficients, signs and significance levels, whereas OLS is more suitable than Poisson due to the predicted values and the adjusted R-squared.

Table 2.1.6: Comparison of OLS and Poisson estimation results

This table compares the average partial effects (APE) of the Poisson regression with the results of the OLS estimation for formal and informal VC.

VARIABLES	Actual data	OLS MaxFon		Poisson MaxFon		OLS MaxFon		Poisson MaxFon	
		(1a)	(2a)	(1b)	(2b)	(3a)	(4a)	(3b)	(4b)
formal VC									
Obs	1,002	1,002	1,002	1,002	1,002	1,002	1,002	1,002	1,002
Mean	1.389	1.387	1.387	1.387	1.387	1.387	1.387	1.387	1.387
Std. Dev.	1.734	1.581	1.585	1.732	1.741	1.589	1.591	1.762	1.762
Min	0	-0.396	-0.466	0.391	0.384	-0.626	-0.570	0.225	0.225
Max	8	7.628	6.850	11.768	12.849	6.639	6.817	13.662	12.954
Adjusted R-squared	1	0.830	0.836	0.703	0.721	0.839	0.841	0.730	0.736
informal VC									
Obs	11584	11584	11584	11584	11584	11584	11584	11584	11584
Mean	1.389	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
Std. Dev.	1.734	1.350	1.350	1.350	1.350	1.354	1.354	1.354	1.354
Min	0	-0.068	-0.522	-0.068	-0.062	-0.342	-0.308	-0.342	-0.308
Max	8	6.964	6.118	6.964	6.170	6.063	6.345	6.063	6.345
Adjusted R-squared	1	0.830	0.836	0.805	0.806	0.839	0.841	0.810	0.810

With respect to our discussion in Section 2.1.3.2, we perform another investor-activity-based robustness check, by focusing on active investors only. In line with the previous literature, we consider VCs with less than 3 investments as passive and remove them from the sample. Hereby, we end up with a subsample of 81 fVCs and 238 iVCs. An OLS regression performed with this subsample leads to similar coefficients. Yet, we see a sporadic drop in significance, which is likely happening due to the smaller sample size. In summary, our

[17] This term has been used by the econometrician Jeffrey Wooldridge to describe the phenomenon observed in our analysis. X-values in the lower and middle range of the spectrum result in predicted y-values which are close to the actual values. However, x-values in the upper range of the spectrum result in predicted y-values which exceed the actual ones significantly.

robustness results confirm the validity of our OLS regression analysis with respect to the investment activity of the investors.

2.1.5 Discussion and implications

We examine 14,575 investments from 12,588 investors in 3,328 companies that have taken place within the setting of the entrepreneurial ecosystem of Cambridge (UK) during the period of 2007-2016, to provide the first theoretically based and empirically tested comparison between formal venture capitalists (fVCs) and informal VCs (iVCs) with respect to their investment diversification strategies and respective portfolio returns. Prior literature has stayed silent on whether and how iVCs actively influence their portfolio returns through various investment strategies. We consider two types of diversification strategies, namely investment diversification across different industries and investment diversification across stages (of the entrepreneurial journeys) and we provide novel insights regarding the iVCs investment diversification strategies vis a vis their formal counterparts.

2.1.5.1 Implications for theory and practice

Our paper makes three distinctive contributions to the literature of entrepreneurial finance. First, due to the structural differences between fVCs and iVCs, and in line with our hypotheses, we find that iVCs benefit less from industry and stage diversification than fVCs. More precisely, the positive relationship is more than twice as large for fVCs than for iVCs. In other words, both types of investors benefit from spreading their investments across industries and venture stages, but for iVCs such an approach returns less value than for fVCs. While these results do not validate our theoretical argument for the advantages of specialization in iVC investment strategies (which would have been reflected by an insignificant coefficient on the effects of diversification on our dependent variable), they argue for less diversified (eq. more specialized) portfolios adopted by iVCs. For example, more experienced iVCs in our analysis perform better through less, comparatively, diversification of their investments, a signal that

could be attributed to focusing on certain types of investments where they hold operational experiences. Additionally, our results reflect the limited investment capacity that iVCs have, which might push them to contribute to fewer rounds of investments despite the successful choice of ventures to invest. This likely prevents them from leveraging the full advantages of diversification. Nevertheless, it is the comparison between iVCs and fVCs that allows us to draw more refined insights into the iVC strategies.

Second, we qualify how properties of the investment strategies undertaken by the investors (iVCs in particular) shed more light on the exact benefits of diversification. Higher prior investment experience in terms of the number of investments most likely makes both types of investors benefit by focusing on fewer industries. In line with our previous findings, fVCs and iVCs are initially more successful when diversifying across industries, however, increasing investment experience reverses the effect and makes industry specialization increasingly more successful. Said differently, prior investment experience acts as a substitute for the need to diversify across industries. Contrary to our theory motivated hypothesis, the moderating effect of prior investment experience is much stronger for iVCs than for fVCs. This implies that prior investment experience is a stronger substitute for the need to diversify for iVCs. Therefore, iVCs can benefit more from focusing on specific industrial contexts upon the buildup of sizeable investment experience. This way they can better leverage advantages, e.g., an industry-specific reputation or preferred deal flow access. While we find no additional properties that influence the importance of diversification for fVC investment portfolios, we identify that several other properties moderate the value of diversification for iVC performance. We find that iVCs, who on average make investments in more mature ventures, i.e., ventures at *later* stages of their development, benefit more from a more diverse portfolio of these investments. Our finding indicates that at later investment stages, it becomes increasingly difficult for iVCs to gain an industry- or stage-specific competitive advantage through specialization. A final feature that affects the iVCs portfolio diversification impact on performance is the average risk

of the industries included in the portfolio of the investors. The higher this average risk of the industries included in the portfolio is, the less successful diversification (eq. the more successful specialization) becomes. This result leads to an interesting realization: a riskier set of industries increases the systemic risk of the venture portfolio making it likely to fail; naturally, efforts to diversify across more industries bear less fruit in such environments. Hence, the common perception that higher diversification is most successful in higher-risk settings established in the extant investment theory fails to grasp the nuances present in early entrepreneurial investments: as the individual investors (iVCs) operate with a target return in mind and aversion to the downside, then higher average industry risk, due to the industries included in their investments, prompts them to avoid negative outcomes by focusing on fewer industries and stages. Also, the more complex and riskier the industries included in the investment portfolio are, the more effort it takes to fully understand the industry- and stage-specific challenges involved. Hence, investors may limit their "spread" across fewer industries or stages respectively to ensure more attention and capacity to better understand their invested ventures. Once an investor gains the respective understanding, it likely provides a competitive advantage and, thus, the iVC should focus future investment activities on similar settings, i.e., pursue industry and stage specialization rather than diversification.

Third, our results suggest that independent of investor types, the resulting returns are highly skewed, i.e., returns are distributed according to a power-law and they depend on one or few home run investments per portfolio for both fVC and iVC. Despite similar return profiles, we observe that iVCs tend to minimize their downside risk rather than maximizing their upside potential, whereas fVCs focus on maximizing their upside potential without too much emphasis on protecting their downside.

2.1.5.2 Limitations and suggestions for future research

We are well aware that similar to all research efforts, our work is conditioned by several limitations. To that effect, we have consistently tried to ensure the robustness and validity of

our research with respect to the following issues.

First, we argue that *MaxFon* is a meaningful approximation of a VC's internal rate of return (IRR). Although the dependent variable construction process is logically concise and verified through robustness tests, it is still not as precise as financial metrics such as IRR and might misrepresent some outlier cases. Our approximation seems to be suitable for this first effort to relatively compare fVC and iVC investment strategies; still, there might exist more suitable metrics or approximations for investor performance and, in that light, we believe that future efforts should try to explore them with newer approaches.

Second, there may exist endogeneity effects between the available capital per investor as well as the previous operational experience per investor and the pursued strategy. While the differences in capital availability for iVCs and fVCs, among other reasons, explain the distinct magnitudes of the diversification-performance relationships, a similar effect might exist within both investor groups which in turn may imply unobserved effects on investment strategy choices. Similarly, there may exist unobserved effects with respect to the investor's previous operational experience. In other words, the natural benefits of specialization might oppose the positive effects of diversification less for iVCs with higher capital availability (less operational experience) than for those with lower capital availability (more operational experience). Unfortunately, our data cannot allow for the level of granularity required to address fully the issue of differences in capital availability and previous operational experience per investor. The possibility to seek further detailed metrics remains an open question for future research.

Last, with respect to our bottom-up data collection approach and the comprehensiveness of the resulting dataset, we rely on the assumption that both the fVCs and the iVCs invest *en masse* within the focal ecosystem of our study, i.e., the Cambridge entrepreneurial cluster (aka Silicon Fen). Our assumption is most likely true for the majority of iVCs, and several anecdotal discussions with different angel investors support our expectation. However, we likely miss a non-negligible proportion of fVC investment, who may be active across ecosystems. Still, the

fact that our findings are in line with the previous literature on fVC confirms to a comfortable extent the validity of our approach. Given that our major contribution lies with identifying robustly the iVC strategies – than on reiterating findings on fVC investments – we still hope that future research will explore alternative data collection approaches that will aim to holistically represent all portfolios.

Our research tries to answer important questions regarding the early-stage investment strategies of venture capitalists with a special emphasis on informal VCs, but naturally, it cannot exhaust the many important issues present in entrepreneurial finance. Nevertheless, we hope that our findings add a valuable building block in the broader effort to better understand early-stage investing, and particularly the underlying conditions that shape the entrepreneurial investors' eventual success.

2.2 Essay 2 – Benchmarking Venture Capital Databases

Abstract

There has been an increasing asymmetry between the rising interest in private companies and the limited availability of data. While a group of new commercial data providers has identified this gap as a promising business opportunity, and has started to provide structured information on private companies and their investors, little is known about the quality of the data they provide. In this paper, we compare detailed and verified proprietary information on 339 actual venture capital (VC) financing rounds from 396 investors in 108 different (mostly European) companies, with data included in eight frequently used VC databases to help academic scholars and investors better understand the coverage and quality of these datasets and, thus, interpret the results more accurately. We find that greater financing rounds are more likely to be reported than lower ones. Similarly, financing round sizes and post-money valuations are more likely to be reported for greater financing rounds than for lower ones. Our analysis reveals that VentureSource, Pitchbook and Crunchbase have the best coverage, and are the most accurate databases across our key dimensions of general company data, founders and funding information. We describe our findings in detail and discuss potential implications for researchers and practitioners.

Keywords: Venture capital, startups, database, benchmarking

Authors: Andre Retterath, Reiner Braun

First Author: Andre Retterath

Current Status: Working paper

2.2.1 Introduction

With the rise of venture dollars invested (Atomico, 2019; Statista, 2019), companies are increasingly able to stay private longer, which in turn shifts significant parts of the value participation from public to private investors. Ritter (2015) finds that the average US tech company that went public in 1999 took about 4 years to do so (from establishment to becoming a publicly-traded and owned entity), whereas in 2014 the average was 11 years. Our own analysis of European tech companies reveals that companies which went public in 1999 took on average 7 years, whereas in 2019 they took on average 12 years from incorporation. In line with Ritter's findings on US tech companies, we also find that the proportion of European tech companies achieving a valuation of more than ten billion dollars in private status jumped from 3% between 2000 and 2009 to 21% between 2010 and 2019[18]. This is evidence that interest in private companies has gained strong momentum. As academic researchers seek to understand a wide spectrum of questions related to topics such as investment strategies, entrepreneurial behavior, social capital, economic impact, and innovation more generally, investors in private firms rely on such sources for several aspects of their work, e.g., when collecting information on potential investment targets such as previous financing rounds, existing shareholders and team backgrounds.

However, it is widely known that private companies – in particular early on in their existence – are surrounded by severe information asymmetries. Entrepreneurs cannot be expected to be accurate in providing information because they have obvious benefits from exaggerating firm quality (Brealey et al., 1977). In turn, investors have a hard time identifying high-quality firms and need to obtain additional, reliable information. However, producing reliable information about small private companies in such an imperfect market is very costly for an individual investor (compared to the situation for public stock markets.) Hence, the

[18] We base our analysis on data provided by Pitchbook and cross-checked via Crunchbase and VentureSource.

incentives for financial intermediaries to engage in such production are comparably weak. In recent years, however, digitization and automatization have enabled large-scale data collection and have gradually reduced the associated cost of such information production. Several service providers have identified this opportunity and started to leverage technology to serve the gap between the increasing need for private company information and the limited availability of data. It is particularly attractive as these independent database providers collect information just once, but can distribute it (in theory) to an infinite number of customers. While technology has obviously increased data availability, little is known about the quality of data that is provided by such commercial databases and the biases which underlie them. In addition, it may well be that they provide superior information for firms with true firm values above the average. Such entrepreneurs have an incentive to send relevant signals to acquirers of their shares (picked-up by databases) because it will result in an increase in share price (Ramakrishnan and Thakor, 1984). Hence, VC databases should be more likely to collect self-reported information on higher-quality firms, resulting in a positive selection bias.

Although Kaplan and Lerner (2016), Kaplan, Strömberg, and Sensoy (2002), Lerner (1995), as well as Maats, Metrick, Yasuda, Hinkes, and Vershovsk (2011), have revealed that even some of the most established VC databases were inconsistent and incomplete, their question of "*How well do VC databases reflect actual investments?*" has neither been applied to today's prevalent databases nor extended by non-transactional data, which becomes increasingly more important. In Kaplan and Lerner's (2016) words: "*While many of these newer databases are promising, they have not gotten the kind of scrutiny that VentureSource and VentureXpert have. Thus, their ability to support academic research is still to be fully determined.*"

The objective of this paper is to shed more light on these newer databases, and to understand the extent to which they are complete (data quantity/coverage) and to which they correctly represent information (data quality/accuracy) which is crucial for researchers and

practitioners alike. We aim to extend previous efforts by including 'newcomer' VC databases not covered before, and which go beyond transactional data. We compare proprietary actual contracts and investment documentation of 339 VC financing rounds from 396 investors in 108 different, mostly European, companies with their characterisation in the eight most relevant VC databases across three primary dimensions: (1) general company, (2) founders and (3) financing information. The data is sourced from ten European VC partnerships that invest globally.

We determine the most relevant VC databases[19] among academics by searching the relevant empirical VC literature and counting the frequencies with which the databases are used. In terms of practitioners' usage, we ran a survey of 111 European VC firms. These two exercises resulted in a shortlist of the most frequently cited and used databases comprising of Angellist (AL), CB-Insights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

Our analysis covers a wide range of variables and details. On average, VS, PB and CB seem to have the best coverage and are the most accurate databases across all the relevant dimensions. VS consistently has the best coverage and quality for the analyzed general company information. PB provides the best coverage and quality for all founder-related information. Concerning funding information, CB has the best coverage with respect to financing rounds and total capital committed, whereas VS and PB have the best coverage and accuracy in terms of round sizes and post-money valuations, respectively. Consequently, a combined dataset with the best possible coverage would consist of general company information from VS, founder information from PB and funding information from a combination of CB, PB and VS. In line with arguments of entrepreneurial signaling or relative cost of information production, we find that greater financing rounds are more likely to be reported than smaller ones. Similarly, financing round sizes and post-money valuations are

[19] In line with the previous literature (Kaplan and Lerner, 2016; Kaplan, Strömberg, and Sensoy, 2002; Lerner, 1995; Maats, Metrick, Yasuda, Hinkes, and Vershovsk, 2011) we refer to these private company databases as "VC databases."

more likely to be reported for greater financing rounds than for lower-value ones. Although our results reveal a variety of further biases, it is hard to summarize them into consistent patterns across all databases. As a consequence, we propose a solid understanding and consideration of all nuances as described in Section 2.2.4, as it might otherwise materially impact any kind of research results.

Clearly, our analysis is conditioned by limited generalizability, as it is focused on a specific subset of early-stage ventures and was conducted at a fixed point in time. Our results might vary for companies in a different development stage, geography or industry, but also for datasets collected at another point in time. Nevertheless, we believe that our study helps scholars and practitioners to better understand the coverage and biases of their data and interpret the results more accurately. Such validations seem particularly important to us in the light of recent efforts to apply machine-learning methods to assist VC decision-making, e.g., in the investment selection process. To deliver meaningful results, such methods require unbiased training samples, and little is known as to which databases are expedient.

The remainder of this paper is structured as follows: In Section 2.2.2, we identify the currently most relevant VC databases, and review academic studies that have used their datasets. We cluster these studies into groups based on their field of research, and identify frequently considered information. Subsequently, Section 2.2.3 describes our method and compares the reported data with the actual information. Section 2.2.4 presents multivariate regression results and describes the determinants of inclusion. Lastly, Section 2.2.5 discusses the potential implications and limitations of our study.

2.2.2 Most frequently used databases, applications and key variables

In this chapter, we follow a top-down structure to comprehensively reveal the use of VC databases, and to create the basis for our further analysis. We count the number of academic articles published based on a specific VC database over the last ten years and compare it with

the penetration analysis of our investor survey. Following the same structure, we dive into more detail by describing potential research questions approached by academics and practical applications for investors, all based on the identified databases. Subsequently, we aggregate the multitude of variables utilized and cluster them into three distinct groups. Only this level of detail enables us to grasp the impact of potential biases and put our subsequent findings into context. Lastly, we summarize previous benchmarking studies and their results.

2.2.2.1 Identifying databases for our benchmarking

We follow a dual approach to identify the most frequently used databases for both academics and practitioners. Concerning academia, we pursue a combination of a top-down and bottom-up approach. Firstly, we search Google Scholar as of December 31st, 2019 with all possible combinations of the keyword group {"venture capital," "VC," "startup"} and the keyword group {"database," "data"}, and manually select those articles published between January 1st, 2009, and December 31st, 2019, independent of the respective journals. We thus identify seven major VC databases, namely AL, CB, CI, DR, PB, PQ and VS. Secondly, we reverse the initial step and screen all papers that have been published based on the previously identified databases and within the same period of time to quantify the respective annual database penetration. We substitute the first keyword group with the names of the identified databases, and search Google Scholar for the respective combinations such as "Crunchbase database" or "Pitchbook data." This top-down-bottom-up approach results in a total of 690 academic papers as exhibited in Table 2.2.1.

Table 2.2.1: VC database penetration for academics and practitioners

Overview of the VC database penetration in academic papers and across VC practitioners. Through a combination of a top-down and a bottom-up approach, we have identified 690 academic papers through Google Scholar that are based on the identified VC databases and that were published between January 1st, 2009, and December 31st, 2019. The table reports the identified number of papers per database per year together with the % of Total papers per year. On the right, the table describes the number of responses and the respective % of Total of our VC survey from December 2019. Multiple selection was allowed and 111 investors submitted 273 responses. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

	Total	2009		2010		2011		2012		2013		2014		2015		2016		2017		2018		2019		VC Survey Dec 2019	
Database	Reported	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Reported	% of Total	Respondent	% of Total
AL	24	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	2	3%	10	11%	6	5%	2	2%	4	3%	33	12%
CI	38	0	0%	0	0%	0	0%	0	0%	0	0%	1	3%	2	3%	5	6%	13	11%	11	8%	6	4%	35	13%
CB	207	1	5%	2	11%	2	12%	4	17%	6	21%	12	32%	15	25%	26	30%	36	31%	50	38%	53	35%	88	32%
DR	3	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	2	2%	1	1%	0	0%	27	10%
PB	162	1	5%	2	11%	0	0%	3	13%	4	14%	5	13%	6	10%	19	22%	32	28%	35	27%	55	37%	63	23%
PQ	166	2	11%	4	21%	4	24%	12	50%	11	39%	14	37%	23	38%	22	25%	20	17%	28	21%	26	17%	8	3%
TR	0	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	15	5%
VS	90	15	79%	11	58%	11	65%	5	21%	7	25%	6	16%	12	20%	6	7%	7	6%	4	3%	6	4%	4	1%
Total	690	19	100%	19	100%	17	100%	24	100%	28	100%	38	100%	60	100%	88	100%	116	100%	131	100%	150	100%	273	100%

For each database and year, we show the total number of articles found and the percentage proportion of all papers in a given year. In our total sample period, we observe four different market leaders: VS until 2011, PQ between 2012 and 2015, CB between 2016 and 2018, and PB in 2019. VS was the most dominant VC database in earlier years, but its relative usage declined substantially over more recent years. A somewhat similar, but less pronounced pattern can be observed for PQ, which went from market leader (2012 to 2015), with a share of at least around 40%, to below 20% more recently. While the field of relevant databases diverged from one (VS) to many between 2011 and 2016, it seems that the number has been converging again from multiple to few. After 2015, CB and PB became the dominant data sources for academic VC research according to our analysis, together representing between 60-70% of the articles published in these years.

Concerning practitioners, in December 2019 we conducted a survey and collected feedback from 111 European VCs about their database usage. 88 of the respondents are institutional VCs, 4 corporate VCs, 5 family offices active in the asset class, and 14 "others" such as Accelerators or Incubators. The investors were provided with a list of the seven VC databases identified from academic research, and a free input field to add any databases not listed. Multiple selection was allowed. For direct comparability between academic and practitioner penetration, Table 2.2.1 shows the survey results of December 2019 right next to the academic papers in 2019. Multiple respondents mentioned TR as an additional database in the free input field, thus we added it to the table. Additional databases which were mentioned no more than twice in the free input field include "Beauhurst," "LinkedIn Company Search" and "Startup Detector." Due to their similarly limited relevance for academics and VC practitioners, as well as their limited coverage of companies, we decided to exclude these additional databases from our benchmarking. In line with the academic database penetration in 2019, CB and PB have the highest relevance for investors, at 32% and 23% respectively. The following positions are taken by databases that are barely or not at all used by academics: AL,

CI, TR and DR. The dominant databases in academic research from earlier years, PQ and VS, have a market penetration among practitioners in our sample of 5% and 1%, respectively.

According to their own methodological descriptions, all database providers leverage a set of similar automated approaches to collect and validate data, but seem to have different resource capacities for ensuring data quality through human intervention. More database characteristics and data collection methods details are provided in Appendix A. They were gathered from the respective company websites, and from details provided by company representatives. As a result of this comparison, we expect similarly high coverage with respect to companies and investors, but significant differences in terms of data quality.

2.2.2.2 Database applications

As described earlier, some of these databases are more relevant to academic researchers, while others are more important to practitioners. In order to holistically understand the diverse applications and the impact of potential flaws, we first cluster previous academic papers based on the abovementioned databases into their respective scholarly subject areas and look into specific research questions, before describing the different use cases of VC investors across their value chain.

Academic use of VC databases. Most academic papers premised on the above-mentioned databases can be assigned to "Business, Management, Accounting" and "Economics, Econometrics and Finance" with journals like the "Journal of Business Venturing," "Entrepreneurship Theory and Practice," "Journal of Business Research," the three top journals in Finance, but also outlets like the "Journal of Private Equity" or "Journal of Economic Behavior and Organization" respectively. Scholars address questions related to founders' personality (Winkler et al., 2019), venture performance (Croce et al., 2018), investment selection (Thies et al., 2019) or entrepreneurial ecosystems more generally (Achleitner et al., 2019; Bonini and Capizzi, 2019; Braun, Weik, and Achleitner, 2019). Related areas include "Social Sciences" with the "Administrative Science Quarterly" and "Environmental Science"

with the "Journal of Cleaner Production." Scholars examine questions including the social capital of VCs (Ter Wal et al., 2016) or the investor's sustainability (De Lange, 2019). While approaching a diverse set of questions, all researchers share one commonality: they are all aware of potentially incomplete, biased or wrong information provided by the VC databases, and thus tend to complement or verify their data with additional sources. As described earlier, it comes down to a trade-off between scalability/sample size of the dataset and trustworthy/unbiased results.

Another well represented and strongly growing research area is "Computer Science" with journals such as "IEEE," "Information Services and Use" and "Empirical Software Engineering." In this discipline, researchers increasingly leverage VC datasets to train predictive models and identify the most promising investment opportunities based on intelligent models (Arroyo et al., 2019; Krishna et al., 2016). Unfortunately, these researchers lack clarity in terms of the comprehensiveness and quality of their training data, and their results therefore need to be taken with a grain of salt. As a consequence, these unknown unknowns result in a lack of trust in such models, and prevent the real-world application of the latter.

Clearly, it would be helpful for researchers across all disciplines to better understand the comprehensiveness, quality and potential biases of these datasets in order to interpret their findings more accurately, and to create trust and credibility vis-à-vis third parties.

Practitioner use of VC databases. To comprehensively understand the importance of such databases for VC practitioners, we disentangle the VC value chain and sequentially analyze the respective usage. There are two widely accepted classifications for the investment process, i.e., a five-staged process (Tyebjee and Bruno, 1984) and a six-staged process (Fried and Hisrich, 1994). The former distinguishes between pre- and post-investment process, whereas the latter differentiates the screening and evaluation stage more granularly. We merge both frameworks and aggregate the respective stages based on practitioners' use of VC databases.

Sourcing. This describes the initial phase of the funnel in which investors identify potential targets. The goal is to identify promising opportunities as early as possible, so as to get into pole position and win the best deals. The deal flow, i.e., new investment opportunities, is clustered into inbound and outbound channels. Inbound deal flow describes those opportunities which directly or indirectly approach the investor via a variety of channels, whereas "outbound" describes the deal flow which is actively identified and approached by the investor. As investors can only indirectly impact the inbound deal flow through branding or networking activities, data-driven approaches naturally apply to the outbound channel. Within the data-driven outbound channel, we classify by inhouse and outhouse solutions. Inhouse solutions mainly involve web crawlers which collect information from app stores, product hunt websites, Git repositories, public registers, and accelerator websites, among others. Furthermore, manual research activities and deep dives conducted by the investment team are considered as important sources of inhouse outbound activities. Outhouse outbound solutions comprise external matchmaking services such as Aingel.ai, Capital Pilot or Crunchdex, as well as external databases, which provide tailored investment opportunities based on the investor's filters. From anecdotal evidence, we are very confident that the use of external VC databases is the prevalent and most important source of outhouse outbound deal flows. High coverage is of the utmost importance, as VCs cannot afford to miss out on promising investment opportunities. Thus, the better the coverage of startups in terms of geography, industry and stage of interest, the higher the value for the investor.

Screening and evaluation. VCs have limited resources both in terms of time and capital. Consequently, they need to identify those companies with the highest likelihood of success, and narrow the deal-flow funnel as efficiently and as fast as possible. Historically, this has been done manually by collecting multiple datapoints on the

ventures of interest through desk research or expert discussions, and ultimately relying on a combination of gut feeling, heuristics and experience. Similarly to sourcing, however, VCs have gradually started to increase the degree of automation in their selection and due diligence process, and started to leverage different data-driven approaches. They can be dissected into inhouse and outhouse solutions. Inhouse selection approaches comprise a variety of machine learning models, similar to the ones pursued by academics (Arroyo et al., 2019), and deterministic algorithms to spot direct or indirect success metrics such as employee growth, news mentions, product ratings and social media activity. Outhouse selection approaches involve predictive success scores for the underlying companies, increasingly provided by a variety of database providers. Examples include the "Minicorn Score" by TR and the "MOSAIC Score" by CI. Furthermore, these databases provide competitive intelligence to automatically identify and compare potential competitors of the company of interest. Market intelligence is another dimension in which databases provide information on market sizing, growth analysis and market fragmentations. From anecdotal evidence, we assume that investors collect multiple datapoints to form their initial perspective and decide which teams to subsequently spend their time with and to invest in. Wrong information likely misleads an investor's decision and can have a negative impact on the ultimate fund performance. Clearly, the quality of the respective VC databases is of the utmost importance in the selection and due diligence process.

Post-investment activities: The phase after the sourcing, screening, evaluation and ultimate investment is the portfolio work. In this phase, investors serve as board members or advisors, and seek to add as much value as possible to accelerate the company's performance. Though portfolio support is mainly manual and spans a variety of tasks, it can be complemented through external data such as competitive intelligence or market analytics as described above. Moreover, these databases can help to identify

suitable follow-on investors through investment-activity-based filters. Based on anecdotal evidence, we assume that the use of external VC databases and analytics is less important and less common than in the pre-investment stages.

Fundraising. Although not included in the investment process by Fried and Hisrich (1994) or Tyebjee and Bruno (1984), institutional VCs regularly need to raise funds from external investors, i.e., their LPs. As previously described, VCs spot their investment opportunities through a variety of channels, one of them being external databases. Similarly, LPs spot their investment opportunities, i.e., VC funds, through a mix of inbound channels, like placement agents or direct outreach from VCs, and outbound channels driven by external VC databases and manual market research. Besides their own sourcing process, LPs leverage these databases to collect information on their potential VC targets and narrow their funnel. The same logic with respect to the relationship between VCs and startups applies to LPs and VCs, just on another level. While it is in the best interest of startups to be correctly represented in these databases in order to be searchable by VCs, it is in the best interest of VCs to have their information correctly represented in order to be searchable by LPs. Coverage and quality are similarly important. Diving more deeply into the above-mentioned databases, we find that some providers such as PQ and DR focus more on the VC-LP relationship than others such as AL or CB. Besides comprehensive startup information, the former databases provide information on VC performance and fund-level metrics.

In conclusion, we find that VC practitioners leverage external databases across their value chain and put varying emphasis on data coverage and quality. Next, we compile the most relevant variables used by academics and practitioners to better understand the impact of missing or wrong information.

2.2.2.3 Most frequently used variables

Although there is a wide range of questions that can be approached via the VC databases, the most relevant and frequently used information for such analyses can be clustered into three distinct groups (Arroyo et al., 2019).

Company information. This dimension spans all static information at a company level which, unless there has been a pivot, does not change over time. It includes the founding year, location, industry classification, product description or information on the business model.

Founders. This category includes all team-related information and is based at a company level. Variables of interest include the founders' gender, age, education (highest level of qualification, subject area of degrees, year of graduation, universities), previous experience (industry-specific, leadership-specific, startup versus corporate), as well as social media activities and social network connections.

Funding. The variables in this group include all financing-related information across two levels. At the company level, it includes the ownership structure (capitalization table), the total amount of capital invested, the total number of financing rounds, the total number of investors, the most recent financing stage, as well as the recent valuation. At the deal level, it includes information such as the size, the valuation, and the investors participating in each financing round.

2.2.2.4 Previous findings

There have been few attempts made to examine the completeness of VC databases. One of the initial approaches by Lerner (1995) analyzes ThomsonOne, formerly known as VentureXpert and as Venture Economics, and shows that the number of VC financing rounds is overstated due to staged investments which were reported as multiple stand-alone financing rounds. Kaplan et al. (2002) compare proprietary information for 143 financing rounds between 1986 and 1999 in 98 companies obtained from fourteen VC partnerships with their representation in VS, formerly known as VentureOne, and ThomsonOne. They find that both

databases exclude about 15% of the actual financing rounds and 20% of the capital committed. The coverage significantly drops with respect to post-money valuations. While VentureSource misses 30% of the valuations, ThomsonOne misses roughly 70. Additionally, the authors analyze both databases towards sampling errors with respect to specific geographies, stages, industries, round sizes and dates of the financing rounds. Indeed, they find that ThomsonOne and VS oversample larger rounds and California-based companies. Although, VS has a better coverage of valuations, it is biased towards reporting post-money valuations of companies with higher valuations. Maats et al. (2011) iterate this approach by comparing the actual investment data of 449 venture backed companies with their representation in VS and ThomsonOne. Kaplan and Lerner (2016) summarize previous VC database studies and highlight the inherent challenges and consequences of potential biases and misrepresented or omitted information, before they translate the question from a startup financing context to a VC fund performance level. They find that Burgiss likely provides the best performance data on VC funds. Besides their specific findings on the VC-LP relationship, and VC performance more specifically, this study exemplifies the variety of applications and the multidimensional character of such databases.

In summary, previous studies have focused on two different units of analysis: the investee/company (Kaplan et al., 2002; Lerner, 1995; Maats et al., 2011) and the investor/VC fund (Kaplan and Lerner, 2016). With respect to the variables of interest, however, researchers have purely focused on one of the three dimensions, i.e., funding information, and completely ignored the remaining two dimensions of team and general company information. Besides the unaddressed variables of interest, Kaplan and Lerner (2016) have identified an increasing gap between the growing use of novel VC databases and the lack of available insight into their coverage and quality: "*While many of these newer databases are promising, they have not gotten the kind of scrutiny that VentureSource and VentureXpert have. Thus, their ability to support academic research is still to be fully determined.*" Although Dalle, Den Besten, and

Menon (2017) provide a comprehensive description of the usage of CB, they have been unable to analyse the coverage and quality of the database itself. The remainder of this paper attempts to fill the above-mentioned gap across the eight most relevant VC databases.

2.2.3 Comparative analysis

2.2.3.1 Actual sample

We asked ten European VC partnerships to provide original documents regarding all of their investments. The documents contain detailed information on the financing round size, round structure, pre- and post-money valuation, fully diluted ownership as well as general company information such as headquarters, industry, founding year, previous investors, and founder backgrounds, among others. To ensure that we can be absolutely confident of the correctness of the original data, we only consider those companies in our sample for which we have all information and the complete financing documents. We thus exclude 8 companies and end up with a proprietary dataset of 108 portfolio companies that received 339 financing rounds from 396 globally active VC partnerships between January 1, 1999 and July 1, 2019.

The first column ("actual") in Table 2.2.2 describes the original sample across the general company dimensions of geography and industry. 56 companies (52%) are based in Germany, 13 companies (12%) in the US and 9 companies (8%) are headquartered in Turkey. The rest (28%) are almost equally distributed across 15 European countries. Concerning industries, 76 companies (70%) are classified as IT/Software companies, whereas 17 companies (16%) are specified as Biotech/Medical/Healthcare, and 15 (14%) as "Others."

Table 2.2.3 depicts the founder-specific education dimensions and the number of founders per company across the actual sample and the respective databases. From the first column ("actual"), we conclude that founders in our sample are highly educated, with 169 out of 296 (57%) having a *Diplom* (equivalent to a master's), a master's degree, MBA or PhD as their highest degree. In terms of the field of their degree, 114 founders (39%) are graduates in

management and 82 (28%) in science, technology, engineering and mathematics (STEM). With respect to the number of founders per company, we find that 45 companies (42%) were started by two founders, whereas only 7 (12%) of them by single founders and 33 (31%) by teams of three. The remaining 21% of companies were founded by teams of four to six members.

Table 2.2.4 describes funding-specific dimensions. The general information is split into "Reported" and "Matched" because some databases report rounds related to the companies that do not exist or cannot be matched. The subsequent analysis of financing rounds per year, financing rounds per company, number of VC investors per company and number of months between rounds are solely based on matched rounds. The first column (actual) shows that 108 companies raised 339 financing rounds with a total value of € 3.442 billion, provided by a total of 396 VC investors. We find that the number of financing rounds over time follows a U-shape, with a local maximum of 18 rounds per year in 2000, a global minimum of 6 rounds per year between 2004 and 2006, and a global maximum of 39 financing rounds in 2018, the last full year in the dataset. The number of financing rounds more than tripled from 11 in 2012 to 39 in 2018. Approximately half of the companies received one or two financing rounds, with 27 and 26 companies, respectively. Only 10 companies received more than 6 financing rounds. The number of VC investors per company across its lifetime approximates a right-skewed normal distribution with approximately two thirds of the companies having two, three or four VC investors.

Table 2.2.2: Sample summary - Company

Summary statistics for actual company information and company information reported by the different VC databases for 108 companies which received financing between 1999 and 2019 . The "Actual" column provides the original information extracted from the available financing contracts and investment memoranda. The "Reported" column represents the information provided by the respective database. The "Rep./Act. %" column presents the ratio of the reported information compared to the actual information. It exhibits the coverage of the respective database in a specific dimension. The "Weighted AVG" line summarizes the weighted average coverage based on the number of actual companies. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

	Actual	AL		CB		CI		DR		PB		PQ		TR		VS	
		Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %
Number of companies	108	48	44%	91	84%	97	90%	89	82%	98	91%	103	95%	91	84%	108	100%
Company location																	
Belgium	1	1	100%	1	100%	2	200%	2	200%	2	200%	0	0%	1	100%	1	100%
Bulgaria	1	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Czech Republic	1	0	0%	1	100%	1	100%	0	0%	1	100%	1	100%	0	0%	1	100%
Finland	2	0	0%	2	100%	2	100%	2	100%	1	50%	1	50%	2	100%	2	100%
France	3	0	0%	3	100%	3	100%	2	67%	3	100%	3	100%	3	100%	3	100%
Germany	56	27	48%	45	80%	45	80%	48	86%	54	96%	55	98%	51	91%	61	109%
Ireland	2	1	50%	1	50%	2	100%	2	100%	2	100%	1	50%	2	100%	1	50%
Israel	1	0	0%	1	100%	0	0%	0	0%	1	100%	2	200%	0	0%	1	100%
Italy	1	0	0%	1	100%	1	100%	1	100%	1	100%	1	100%	0	0%	0	0%
Liechtenstein	1	0	0%	1	100%	1	100%	1	100%	1	100%	0	0%	0	0%	1	100%
Netherlands	1	0	0%	1	100%	1	100%	1	100%	1	100%	1	100%	1	100%	1	100%
Romania	3	0	0%	1	33%	1	33%	2	67%	1	33%	2	67%	2	67%	1	33%
Slovakia	2	0	0%	2	100%	1	50%	1	50%	2	100%	0	0%	2	100%	1	50%
Sweden	2	0	0%	3	150%	1	50%	1	50%	2	100%	2	100%	1	50%	2	100%
Switzerland	5	1	20%	2	40%	3	60%	4	80%	2	40%	3	60%	4	80%	4	80%
Turkey	9	4	44%	5	56%	5	56%	6	67%	5	56%	4	44%	5	56%	5	56%
United Kingdom	4	3	75%	3	75%	3	75%	3	75%	3	75%	5	125%	3	75%	2	50%
USA	13	11	85%	18	138%	25	192%	13	100%	16	123%	22	169%	14	108%	21	162%
Weighted AVG			*44%*		*84%*		*90%*		*82%*		*91%*		*95%*		*84%*		*100%*
Industry																	
Biotech/Medical/Healthcare	17	2	12%	18	106%	17	100%	17	100%	20	118%	12	71%	16	94%	17	100%
IT/Software	76	5	7%	37	49%	62	82%	22	29%	46	61%	24	32%	34	45%	42	55%
Other	15	41	273%	36	240%	18	120%	50	333%	32	213%	67	447%	38	253%	49	327%
Weighted AVG			*44%*		*84%*		*90%*		*82%*		*91%*		*95%*		*81%*		*100%*

Table 2.2.3: Sample summary - Founders

Summary statistics for actual founders information and founders information reported by the different VC databases for 108 companies which received financing between 1999 and 2019. The "Actual" column provides the original information extracted from the available financing contracts and investment memoranda. The "Reported" column represents the information provided by the respective database. The "Rep./Act. %" column presents the ratio of the reported information compared to the actual information. It exhibits the coverage of the respective database in a specific dimension. The "Weighted AVG" line summarizes the weighted average coverage based on the number of actual companies. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

	Actual	AL		CB		CI		DR		PB		PQ		TR		VS	
		Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %
Number of founders	301	65	22%	179	59%	0	0%	126	42%	184	61%	0	0%	142	47%	28	9%
Founders' highest qualification and field of study																	
Bachelor	46	4	9%	18	39%	0	0%	6	13%	21	46%	0	0%	7	15%	0	0%
Master	85	4	5%	21	25%	0	0%	10	12%	35	41%	0	0%	8	9%	0	0%
Diplom	24	1	4%	14	58%	0	0%	0	0%	10	42%	0	0%	7	29%	0	0%
MBA	32	2	6%	17	53%	0	0%	10	31%	15	47%	0	0%	10	31%	0	0%
PhD	33	3	9%	22	67%	0	0%	7	21%	27	82%	0	0%	58	176%	0	0%
Non	1	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	2	200%	0	0%
No Information	80	51	64%	87	109%	0	0%	93	116%	76	95%	0	0%	50	63%	28	35%
Management	114	6	5%	46	40%	0	0%	24	21%	58	51%	0	0%	19	17%	0	0%
STEM	87	7	8%	29	33%	0	0%	17	20%	50	57%	0	0%	2	2%	0	0%
Others	25	2	8%	10	40%	0	0%	3	12%	16	64%	0	0%	2	8%	0	0%
Undefined	19	0	0%	7	37%	0	0%	0	0%	0	0%	0	0%	67	353%	0	0%
No Information	56	50	89%	87	155%	0	0%	82	146%	60	107%	0	0%	52	93%	28	50%
Weighted AVG			*22%*		*59%*		*0%*		*42%*		*61%*		*0%*		*47%*		*9%*
Number of founders per company																	
1	11	17	155%	23	209%	0	0%	20	182%	21	191%	0	0%	19	173%	10	91%
2	77	11	14%	27	35%	0	0%	21	27%	27	35%	0	0%	32	42%	4	5%
3	15	4	27%	14	93%	0	0%	10	67%	21	140%	0	0%	11	73%	2	13%
4	15	2	13%	11	73%	0	0%	2	13%	5	33%	0	0%	4	27%	1	7%
5	5	0	0%	2	40%	0	0%	4	80%	4	80%	0	0%	2	40%	0	0%
6	1	1	100%	1	100%	0	0%	1	100%	1	100%	0	0%	0	0%	0	0%
N/A		13		13		97		31		19		0		23		91	
Weighted AVG			*28%*		*63%*		*0%*		*47%*		*64%*		*0%*		*55%*		*14%*

Table 2.2.4-1: Sample summary - Funding

Summary statistics for actual funding information and funding information reported by the different VC databases for 339 financing rounds in 108 companies from 1999 to 2019. The "Actual" column provides the original information extracted from the available financing contracts and investment memoranda. The "Reported" column represents the information provided by the respective database. The "Rep./Act. %" column presents the ratio of the reported information compared to the actual information. It exhibits the coverage of the respective database in a specific dimension. The differences between reported rounds and actual rounds in terms of round date, round size, post-money and total committed are represented by two blocks "Reported" and "Matched". "Reported" summarizes every round which has been recorded for the listed company and "Matched" summarizes only rounds which have been matched with actual rounds. The line "AVG" presents the average coverage per block. The "Weighted AVG" line summarizes the weighted average coverage based on the number of actual financing rounds. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

	Actual	AL		CB		CI		DR		PB		PQ		TR		VS	
		Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %	Reported	Rep./Act. %
Panel A: Reported																	
Number of rounds	339	83	24%	308	91%	295	87%	296	87%	400	118%	193	57%	245	72%	412	122%
Round sizes	339	75	22%	229	68%	222	65%	208	61%	272	80%	127	37%	188	55%	328	97%
Post-money	339	1	0%	0	0%	59	17%	76	22%	84	25%	0	0%	0	0%	248	73%
Total committed (€ MM)	3442	975	28%	3240	94%	3227	94%	3733	108%	5093	148%	1239	36%	2881	84%	3206	93%
AVG			*19%*		*63%*		*66%*		*70%*		*93%*		*33%*		*53%*		*96%*
Panel B: Matched																	
Number of rounds	339	63	19%	208	61%	186	55%	182	54%	229	68%	167	49%	173	51%	229	68%
Round sizes	339	58	17%	168	50%	156	46%	147	43%	173	51%	114	34%	144	42%	194	57%
Post-money	339	1	0%	0	0%	46	14%	56	17%	50	15%	0	0%	0	0%	169	50%
Total committed (€ MM)	3442	847	25%	2741	80%	2680	78%	2555	74%	2676	78%	1138	33%	2477	72%	2443	71%
AVG			*15%*		*48%*		*48%*		*47%*		*53%*		*29%*		*41%*		*61%*
Total number of investors	396	214	54%	459	116%	471	119%	347	88%	582	147%	424	107%	550	139%	544	137%
Panel C: Year of matched financing rounds																	
1999	12	0	0%	6	50%	1	8%	4	33%	6	50%	7	58%	4	33%	10	83%
2000	18	0	0%	6	33%	7	39%	3	17%	8	44%	8	44%	2	11%	18	100%
2001	13	0	0%	1	8%	3	23%	1	8%	5	38%	7	54%	2	15%	8	62%
2002	9	0	0%	4	44%	1	11%	3	33%	5	56%	2	22%	1	11%	7	78%
2003	8	0	0%	4	50%	1	13%	4	50%	6	75%	2	25%	4	50%	6	75%
2004	6	0	0%	4	67%	3	50%	4	67%	4	67%	4	67%	3	50%	3	50%
2005	6	0	0%	4	67%	2	33%	3	50%	3	50%	3	50%	2	33%	8	133%
2006	6	0	0%	3	50%	2	33%	2	33%	3	50%	2	33%	4	67%	5	83%
2007	10	1	10%	6	60%	3	30%	4	40%	4	40%	4	40%	2	20%	5	50%
2008	7	2	29%	5	71%	5	71%	4	57%	6	86%	7	100%	6	86%	7	100%
2009	15	5	33%	9	60%	6	40%	8	53%	8	53%	11	73%	7	47%	10	67%
2010	15	3	20%	7	47%	9	60%	6	40%	9	60%	9	60%	8	53%	11	73%
2011	15	8	53%	11	73%	9	60%	11	73%	11	73%	9	60%	10	67%	9	60%
2012	11	3	27%	7	64%	6	55%	5	45%	7	64%	5	45%	5	45%	8	73%
2013	18	4	22%	15	83%	13	72%	12	67%	16	89%	11	61%	14	78%	12	67%
2014	25	8	32%	12	48%	13	52%	11	44%	13	52%	12	48%	13	52%	11	44%
2015	34	13	38%	27	79%	22	65%	28	82%	27	79%	18	53%	20	59%	23	68%
2016	25	6	24%	20	80%	20	80%	16	64%	22	88%	9	36%	14	56%	18	72%
2017	33	5	15%	22	67%	25	76%	22	67%	24	73%	12	36%	20	61%	19	58%
2018	39	3	8%	21	54%	19	49%	19	49%	27	69%	15	38%	16	41%	18	46%
2019	14	2	14%	14	100%	16	114%	12	86%	15	107%	10	71%	16	114%	13	93%
Weighted AVG			*19%*		*61%*		*55%*		*54%*		*68%*		*49%*		*51%*		*68%*

Table 2.2.4-2: Sample summary - Funding

Summary statistics for actual funding information and funding information reported by the different VC databases for 339 financing rounds in 108 companies from 1999 to 2019. The "Actual" column provides the original information extracted from the available financing contracts and investment memoranda. The "Reported" column represents the information provided by the respective database. The "Rep./Act. %" column presents the ratio of the reported information compared to the actual information. It exhibits the coverage of the respective database in a specific dimension. The differences between reported rounds and actual rounds in terms of round date, round size, post-money and total committed are represented by two blocks "Reported" and "Matched". "Reported" summarizes every round which has been recorded for the listed company and "Matched" summarizes only rounds which have been matched with actual rounds. The line "AVG" presents the average coverage per block. The "Weighted AVG" line summarizes the weighted average coverage based on the number of actual financing rounds. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

Panel D: Number of matched financing rounds per company																	
1	27	21	78%	30	111%	34	126%	29	107%	33	122%	37	137%	39	144%	32	119%
2	26	7	27%	25	96%	21	81%	24	92%	23	88%	13	50%	19	73%	30	115%
3	17	6	35%	11	65%	9	53%	9	53%	11	65%	12	71%	9	53%	12	71%
4	14	0	0%	3	21%	7	50%	4	29%	3	21%	1	7%	3	21%	4	29%
5	8	1	13%	5	63%	5	63%	4	50%	6	75%	2	25%	3	38%	6	75%
6	6	0	0%	3	50%	0	0%	0	0%	2	33%	0	0%	2	33%	1	17%
More	10	0	0%	0	0%	0	0%	1	10%	1	10%	0	0%	0	0%	2	20%
Weighted AVG			*32%*		*71%*		*70%*		*66%*		*73%*		*60%*		*69%*		*81%*

Panel E: Number of VC investors per company																	
1	8	10	125%	8	100%	9	113%	19	238%	9	113%	12	150%	5	63%	8	100%
2	21	9	43%	14	67%	20	95%	15	71%	11	52%	28	133%	12	57%	17	81%
3	23	8	35%	8	35%	10	43%	10	43%	7	30%	14	61%	10	43%	11	48%
4	25	7	28%	14	56%	11	44%	8	32%	17	68%	13	52%	13	52%	16	64%
5	16	3	19%	12	75%	8	50%	10	63%	9	56%	7	44%	15	94%	10	63%
6	8	4	50%	5	63%	8	100%	6	75%	6	75%	7	88%	5	63%	13	163%
More	7	7	100%	30	429%	31	443%	21	300%	39	557%	19	271%	31	443%	37	529%
Weighted AVG			*44%*		*84%*		*90%*		*82%*		*91%*		*93%*		*84%*		*104%*

Panel F: Number of months matching round date differs from actual date	*Number of rounds in AL*	*% of total rounds in AL*	*Number of rounds in CB*	*% of total rounds in CB*	*Number of rounds in CI*	*% of total rounds in CI*	*Number of rounds in DR*	*% of total rounds in DR*	*Number of rounds in PB*	*% of total rounds in PB*	*Number of rounds in PQ*	*% of total rounds in PQ*	*Number of rounds in TR*	*% of total rounds in TR*	*Number of rounds in VS*	*% of total rounds in VS*
0	17	27%	41	20%	21	11%	64	35%	17	7%	31	19%	24	14%	51	22%
1	18	29%	71	34%	65	35%	54	30%	69	30%	57	34%	59	34%	87	38%
2	7	11%	34	16%	39	21%	20	11%	44	19%	29	17%	31	18%	34	15%
3	5	8%	12	6%	17	9%	9	5%	23	10%	14	8%	13	8%	12	5%
4	1	2%	8	4%	11	6%	9	5%	16	7%	9	5%	8	5%	6	3%
5	1	2%	10	5%	14	8%	7	4%	10	4%	5	3%	5	3%	7	3%
6	2	3%	5	2%	5	3%	2	1%	9	4%	2	1%	12	7%	8	3%
More	12	19%	27	13%	14	8%	17	9%	41	18%	20	12%	21	12%	24	10%

2.2.3.2 VC database benchmarking

We collect all available information for the sample of 108 companies across the eight above-mentioned databases. Tables 2.2.2 to 2.2.6 depict the reported results. The columns represent the respective databases as described above, i.e., "AL" for Angellist, "CB" for Crunchbase, "CI" for CB-Insights, "DR" for Dealroom, "PB" for Pitchbook, "PQ" for Preqin, "TR" for Tracxn, and "VS" for Venture Source. There are two columns per database, for which "Reported" portrays the absolute number and "Rep./Act. %" describes the ratio of the reported value by the respective database divided by the absolute value from the original data. Due to the extensive amount of information collected, we solely highlight the most important findings below.

Company information. Table 2.2.2 shows that VS is the only database with full company coverage. PQ follows with 95%, PB with 91% and CI with 90%, whereas AL has the worst company coverage with 44%. We find the same picture for geographic coverage. Companies headquartered in Eastern European countries are heavily underreported with, on average, less than 50% coverage. Companies in the US and Belgium are overreported with, on average, more than 120%. Overreporting occurs mainly due to a company being reported multiple times with slightly different names, e.g., one entry without the company form such as "Company ABC" and one entry with company form such as "Company ABC Ltd." Companies based in Germany, France and the Netherlands are most accurately reported with an average difference from the actual companies of less than 20%. Concerning industry classifications, all databases perform well on identifying "Biotech/Medical/Healthcare" with an average delta from the actual classification of less than 12%, but underreport "IT/Software" with an average delta of more than 50%. As a consequence, they heavily overreport "Others" with an average delta of more than 250%, likely because "Others" is the collecting bucket for all companies which do not fit into "IT/Software" or "Biotech/Medical/Healthcare."

Founders. Table 2.2.3 shows that PQ and CI do not report founder information at all. PB and CB show the highest average coverage, with 61% and 59% respectively. VS, which has high coverage at a company level, only reports 9% of the founders. For those databases which do report more than 40% of the founder information, the likelihood of a founder's education being provided heavily depends on the degree itself. While the likelihood is three times higher for a PhD than for a bachelors, masters or *Diplom*, it is twice as high for MBAs. Concerning the degree areas, all databases underreport across areas, i.e., they lack this information, with an average delta from the actual degree of more than minus 50%. Lastly, Table 2.2.3 reveals an overarching mismatch between the reported number of founders and the actual number of founders per company. There are two effects: a) either the databases report as company founders managers who are not actually founders, which increases the reported number of founders per company, or b) they miss off one or more founders and thus underreport the number of founders per company. The raw data reveal an overlay of both effects. An average value of 167% for companies with one founder across all databases which report founder information indicates that effect b) is prevalent though. PB and CB match founders most accurately to the companies, with a 64% and 63% weighted average respectively.

Funding. Panel A in Table 2.2.4-1 shows that VS and PB overreport the number of financing rounds, at 122% and 118% respectively. At 91%, a delta of 9% to the actual number of financing rounds, CB is the most accurate database with respect to the number of reported financing rounds. Similar to the general company information, AL has the lowest coverage with 24%. VS reports 97% of the round sizes and 73% of the post-money valuations, almost three times as many as PB (at 25%) and DR (at 22%), and is thus the most accurate database across both dimensions. Concerning total capital committed, PB and DR overreport (at 148% and 108%, respectively), whereas CB, CI and VS underreport, at 94%, 94% and 93% respectively, a delta to the actual amount of 7% or less. Consequently,

the latter three databases are the most accurate with respect to total capital committed. An in-depth comparison between the reported and the actual raw data reveals that these databases over- or underreport specific information for a variety of reasons. For example, several financing rounds contain milestones and have a trenched payment schedule. While these kind of financing rounds are considered as one single round in the actual data, they are oftentimes considered as separate rounds in the reported information, resulting in overreporting. Similarly, we find unique financing rounds being reported with exactly the same information, but different dates. A potential reason could be unverified information collected by web crawlers, e.g., a single financing round mentioned in the news at different dates. On the other hand, internal bridge rounds which have not been externally announced are considered as separate financing rounds in the actual data, whereas external databases either report them as single rounds or not even at all – another reason for underreporting. In reality, these effects are layered and distort the analysis.

The fuzzy matching of the reported financing rounds with the actual financing rounds based on the date of the financing round, the size of the financing round and the participating investors of the financing round, however, allows us to remove those effects. As Panel B in Table 2.2.4-1 shows, PB and VS – the databases with the highest reported coverages – have the highest average drop when comparing matched financing rounds to reported financing rounds, at 42% and 35% respectively. PQ, AL and CB have the lowest average drops (of 8%, 15% and 19%), which indicates a high quality of reported funding information. While CB has the most accurate reported coverage with 91%, the matched coverage drops to 61% and becomes second to VS and PB with 68% each. Similarly, VS and PB have the most accurate matched coverage, with 57% and 51% for round sizes. In line with the reported post-money valuations, VS has the highest matched coverage with 50%, again almost three times as many valuations as DR and PB with 17% and 15% respectively. Concerning the total capital committed, CB, CI and PB have the highest matched coverage with 80%, 78%

and 78% respectively. Although the overall percentage levels drop from reported to matched, we find a similar picture across all variables of interest. VS has the smallest delta between reported and actual across unmatched and matched coverage for number of financing rounds, round sizes and post-money valuations, whereas CB and CI have the smallest delta for unmatched and matched coverage of total capital committed.

Furthermore, Panel B in Table 2.2.4-1 describes the VC investor coverage based on matched financing rounds across databases. Surprisingly, all databases except AL and DR heavily overreport, with on average 128%. An in-depth analysis reveals that most databases wrongly classify non-VC investors as VCs and, thus, report more VC investors than actually participated. Some of them do not even provide a classification. Furthermore, our analysis reveals alternative spellings and the like as another major reason for overreporting, e.g., one unique investor "Investor A Venture Capital" is listed multiple times with different renderings, such as "Investor A VC," "I. A. VC," "Inv. A. Venture Capital," etc. Although PQ overreports, it is the most accurate database because it has the smallest delta to the actual number of VC investors with only 7%.

A comparison of the years of the matched financing rounds (Panel C in Table 2.2.4-1) shows that, again, VS and PB have the highest coverage, both with 68%. Although, all databases except VS and PQ were launched in or after the year 2007, there seems to be no correlation between their respective year of incorporation and the years of covered financing rounds. This indicates that all providers backfill their databases.

With respect to the number of matched financing rounds per company (Panel D in Table 2.2.4-2), we find a similar pattern as for the number of founders per company described above. Companies with one financing round are on average 113% overreported, whereas those with two or more are - with one exception of 115% for two rounds for VS - consistently underreported. With a weighted average of 81%, VS has the most comprehensive coverage of financing rounds per company.

Similarly, Panel E in Table 2.2.4-2 reveals that companies with one VC investor are overrepresented with an average of 125% across databases. On the other hand, companies with up to six VC investors are – with exception of 163% for six investors for VS, 133% for two investors for PQ, and 100% for six investors for CI – consistently underreported. However, companies with more than six VC investors are heavily overreported, with on average 384% across all databases. Again, this is due to incorrect investor-type classification and repeat entries due to variances in spelling etc.

Lastly, Panel F in Table 2.2.4-2 exhibits the number of months the matching round dates differs from the actual round dates. Contrary to all other tables, the presented percentage describes the relative percentage of the total rounds, and thus helps to interpret date accuracy. Independent of the total number of reported financing rounds, DR reports the round dates most accurately, with 35% of them matching the actual month of the financing rounds. More broadly, AL, CB, DR, PQ and VS report more than half of the matched financing rounds with zero- or one-month difference only. PB is the least accurate database with respect to the matched financing round date, as it reports almost every fifth financing round with more than six months' difference. An in-depth analysis reveals that there is a structural delay in reporting, in that the dates of the actual financing rounds t1 are before the reported dates t2 in more than 92% of the cases. This fact can be explained by the time lag between the actual financing round and the public announcement, which serves as the main trigger for these databases to report the round. Cases in which the reported date t2 is before the actual date t1 might be explained by backfilling and incorrect secondary information.

Table 2.2.5: Comparisons of the actual financing round amounts to those reported/matched by the VC databases

Comparisons of the actual financing round amounts to those reported/matched by the VC databases for 339 financing rounds in 108 companies from 1999 to 2019. We compute the ratios (Rep./Act. %) of the reported financing round size to the actual financing round size for all matched financing rounds across the VC databases and report the descriptive statistics (median, mean, standard deviation, standard error and sample size) of these ratios together with their frequency distribution. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

	AL	CB	CI	DR	PB	PQ	TR	VS
Panel A	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.
Median	1.37	1.45	1.68	1.32	1.62	1.57	1.55	1.51
Mean	1.10	1.07	1.05	1.08	1.14	1.04	1.08	1.00
Std.Dev.	1.19	2.90	3.29	0.86	2.93	3.53	3.32	2.83
Std.Err.	0.14	0.08	0.08	0.09	0.09	0.10	0.09	0.07
Num. Rounds	58	168	156	147	173	114	144	194

Panel B: Frequency distribution	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*
0 = x = 0.5	2	3%	10	6%	7	4%	7	5%	8	5%	2	2%	7	5%	6	3%
0.5< x = 0.75	4	7%	8	5%	13	8%	7	5%	12	7%	9	8%	8	6%	13	7%
0.75 < x = 0.85	6	10%	16	10%	17	11%	15	10%	15	9%	13	11%	14	10%	17	9%
0.85 < x = 0.95	6	10%	20	12%	20	13%	13	9%	17	10%	10	9%	14	10%	19	10%
0.95 < x = 1.05	9	16%	31	18%	28	18%	31	21%	21	12%	26	23%	30	21%	61	31%
1.05 < x = 1.15	8	14%	15	9%	16	10%	11	7%	22	13%	15	13%	15	10%	12	6%
1.15 < x = 1.25	6	10%	12	7%	7	4%	8	5%	7	4%	5	4%	7	5%	7	4%
1.25 < x = 1.5	2	3%	12	7%	6	4%	14	10%	26	15%	14	12%	13	9%	11	6%
x > 1.5	15	26%	44	26%	42	27%	41	28%	45	26%	20	18%	36	25%	48	25%

Table 2.2.6: Comparisons of the actual post-money valuations to those reported/matched by the VC databases

Comparisons of the actual post-money valuations to those reported/matched by the VC databases for 339 financing rounds in 108 companies from 1999 to 2019. We compute the ratios (Rep./Act. %) of the reported post-money valuation per round to the actual post-money valuation for all matched financing rounds across the VC databases and report the descriptive statistics (median, mean, standard deviation, standard error and sample size) of these ratios together with their frequency distribution. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

	AL	CB	CI	DR	PB	PQ	TR	VS
Panel A	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.
Median	1.00	0,00	1.22	1.54	1.48	0,00	0,00	1.69
Mean	1.00	0,00	1.08	1.38	1.06	0,00	0,00	1.39
Std.Dev.	0.00	0,00	0.62	0.96	1.81	0,00	0,00	1.31
Std.Err.	1.00	0,00	0.17	0.19	0.16	0,00	0,00	0.11
Num. Rounds	1	0	46	56	50	0	0	169

Panel B: Frequency distribution	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*
0 = x = 0.5	0	0%	0	0%	1	2%	7	12%	2	4%	0	0%	0	0%	9	5%
0.5< x = 0.75	0	0%	0	0%	6	13%	4	7%	2	4%	0	0%	0	0%	10	6%
0.75 < x = 0.85	0	0%	0	0%	4	9%	4	7%	0	0%	0	0%	0	0%	8	5%
0.85 < x = 0.95	0	0%	0	0%	6	13%	6	11%	10	18%	0	0%	0	0%	12	7%
0.95 < x = 1.05	1	100%	0	0%	5	11%	4	7%	10	18%	0	0%	0	0%	14	8%
1.05 < x = 1.15	0	0%	0	0%	4	9%	0	0%	10	18%	0	0%	0	0%	12	7%
1.15 < x = 1.25	0	0%	0	0%	1	2%	4	7%	2	4%	0	0%	0	0%	11	7%
1.25 < x = 1.5	0	0%	0	0%	5	11%	3	5%	7	12%	0	0%	0	0%	20	12%
x > 1.5	0	0%	0	0%	13	29%	25	44%	7	12%	0	0%	0	0%	73	43%

Table 2.2.7: Comparisons of the actual total amounts raised per company to those reported/matched by the VC databases

Comparisons of the actual total amounts raised per company to those reported/matched by the VC databases for 339 financing rounds in 108 companies from 1999 to 2019. We compute the ratios (Rep./Act. %) of the reported total amount raised per company to the actual amount raised per company for all companies across the VC databases and report the descriptive statistics (median, mean, standard deviation, standard error and sample size) of these ratios together with their frequency distribution. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

	AL	CB	CI	DR	PB	PQ	TR	VS
Panel A	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.	Rep./Act.
Median	0.93	0.99	1.12	1.11	1.11	0.77	0.97	1.05
Mean	0.71	0.96	0.89	0.97	1.07	0.63	0.81	0.98
Std.Dev.	1.44	0.57	0.97	0.95	0.72	0.72	0.9	0.74
Std.Err.	0.24	0.07	0.11	0.11	0.08	0.09	0.1	0.08
Num. Rounds	35	76	76	71	79	65	75	87

Panel B: Frequency distribution	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*	Num. Rounds	*% of total rounds*
0 = x = 0.5	11	31%	14	18%	13	17%	12	17%	11	14%	23	35%	18	24%	13	15%
0.5< x = 0.75	8	23%	10	13%	14	18%	12	17%	9	11%	13	20%	15	20%	15	17%
0.75 < x = 0.85	3	9%	8	11%	6	8%	7	10%	7	9%	6	9%	7	9%	4	5%
0.85 < x = 0.95	4	11%	6	8%	8	11%	4	6%	8	10%	4	6%	4	5%	7	8%
0.95 < x = 1.05	2	6%	10	13%	7	9%	10	14%	4	5%	5	8%	7	9%	19	22%
1.05 < x = 1.15	1	3%	5	7%	3	4%	6	8%	12	15%	1	2%	3	4%	3	3%
1.15 < x = 1.25	2	6%	5	7%	8	11%	3	4%	3	4%	5	8%	9	12%	7	8%
1.25 < x = 1.5	2	6%	10	13%	6	8%	7	10%	14	18%	5	8%	4	5%	8	9%
x > 1.5	2	6%	8	11%	11	14%	10	14%	11	14%	3	5%	8	11%	11	13%

Tables 2.2.5, 2.2.6 and 2.2.7 study financing amounts, post-money valuations and total amounts raised per company in more detail. They report the descriptive statistics and frequency distributions of the reported versus actual ratios for the respective dimensions of interest. The unit of analysis for Tables 2.2.5 and 2.2.6 is the financing round, whereas Table 2.2.7 is at the company level. In Panel A of Table 2.2.5 we display descriptives on the ratio of financing-round amounts reported in the corresponding database divided by the actual amount. The numbers displayed show that all databases tend to overstate financing volumes. While a relevant number of understated ratios drag mean values down, all averages for all databases are larger than, or at least equal to, one. On average, VS reports financing round sizes with a 100% accuracy, whereas PQ and CB overreport, with 104% and 107% respectively. In Panel B we assign each financing round detected in a given database into a group of financing-amount accuracy. This illustrates the number of financing rounds with amounts close to the real values. Again, with 31% of reported round sizes within a range of 95% to 105% of the actual round sizes, VS exhibits the highest portion of accurate data points across all databases.

As for financing amounts in Table 2.2.5, for each VC database Table 2.2.6 displays the average accuracy of post-money valuations (Panel A) and the frequency of financing rounds by levels of accuracy. Again, we look at the ratio of the reported divided by the actual amount. And once again, the table reveals that all databases with reasonable coverage of valuations for our benchmarking sample consistently overreport post-money valuations, indicated by mean and median ratios greater than one. With mean ratios of 106% and 108% respectively, PB and CI seem to present the average post-money valuations most accurately. On a more detailed level, the frequency distributions in Panel B show that PB is more precise than CI, with 18% versus 11% of the reported valuations being within a range of 95% to 105% of the actual post-money valuations. This impression is substantiated by extending the interval to 15% around the real value: in PB, 56% of all financing rounds

matched fall into this category, while this only applies to 33% of rounds in CI. Altogether, for financing round sizes and valuation, our benchmarking reveals that most amounts displayed in the VC databases are lower than the actual values, but all of them contain some extreme cases of overreporting.

Lastly, Table 2.2.7 compares the actual total amounts raised per company with those reported by the VC databases. In line with the previous tables, we find a left-skewed distribution, with median ratios greater than the mean ratios. In terms of total amount raised, VS and CB only slightly underreport with mean ratios of 98% and 96%, whereas PB overreports with a mean ratio of 107%. The detailed frequency distribution supports these findings, with VS reporting 22% of the companies within a range of 95% to 105% of the actual total amount raised per company. For our sample, VS performs best in terms of financing amounts and valuations.

Ranking. To allow a comprehensive comparison, we provide an overview of data coverage and quality across our dimensions in Table 2.2.8. VS seems to have the best coverage and quality across all subcategories of general company information. Similarly, PB has the best coverage and quality across all founder-related categories. Although VS, generally speaking, seems to provide the best coverage and quality in terms of funding information, we find that CB dominates coverage in terms of rounds reported and total capital committed.

Table 2.2.8: Database benchmarking

Databases ranked across categories/tables from 1 = best to 8 = worst based on the delta between their respective coverage and 100% for Table 2-4 and based on the total percentage of observations beeing within the frequency distribution of $0.95 < x < 1.05$ for Tables 5-7. We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS).

		AL	CB	CI	DR	PB	PQ	TR	VS
Table	*Company*								
2	Company coverage	8	5	4	7	3	2	5	1
2	Company location accuracy	8	5	4	7	3	2	5	1
2	Company industry accuracy	8	5	4	7	3	2	5	1
	Overall Company	*8*	*5*	*4*	*7*	***3***	***2***	*5*	***1***
	Founders								
3	Founders coverage	5	2	7	4	1	7	3	6
3	Founders education accuracy	5	2	7	4	1	7	3	6
3	Founders completeness accuracy	5	2	7	4	1	7	3	6
	Overall Founders	*5*	***2***	*7*	*4*	***1***	*7*	***3***	*6*
	Funding								
4	Round coverage	8	1	2	2	4	7	6	5
4	Round size coverage	8	3	4	5	2	7	6	1
4	Post-money coverage	5	5	4	3	2	5	5	1
4	Total committed coverage	8	1	1	4	6	7	5	3
4	Matched round coverage	8	5	4	3	2	7	6	1
4	Matched round size coverage	8	3	4	5	1	7	6	1
4	Matched post-money coverage	8	3	4	5	2	7	6	1
4	Matched total committed coverage	8	3	3	5	2	7	6	1
4	Investor coverage	7	3	4	2	8	1	6	5
4	Matched round year accuracy	8	3	4	5	1	7	6	1
5	Matched round size accuracy	7	5	5	3	8	2	3	1
6	Matched post-money accuracy	5	5	2	4	1	5	5	3
7	Matched total raised accuracy	7	3	4	2	8	6	4	1
	Overall Ranking	*8*	***2***	***3***	*5*	*4*	*6*	*7*	***1***
	Overall Ranking	*8*	***3***	*4*	*7*	***2***	*5*	*5*	***1***

2.2.4 Determinants of inclusion (or exclusion)

Following our comparative analysis, we seek to examine the determinants of a company, founder, financing round, financing round size and post-money valuation appearing in the VC databases. The purpose of this analysis is to better understand potential biases. For each combination of the above-mentioned variables of interest and the respective database, we estimate a logistic regression. The dependent variable equals one if the respective variable of interest is included in the database, and zero if not. In line with Gompers and Lerner (2000) and Kaplan et al. (2002), the independent variables are the natural logarithm of the actual total amount of capital raised in millions of Euros (*lnTotalRaised*), a count variable to reflect the actual founding year or the actual year of the financing round (*FoundingDate / RoundYear*), an

actual geography indicator for the most frequent countries (Germany (DE), France (FR), United States (US)) and all others as reference category, an actual industry indicator differentiating between the most frequent industries of software/IT (IT), life sciences, bio technology, healthcare and medical (LS) and all others as reference category, an actual M&A indicator (equal to one if the company has been acquired via a trade-sale / M&A and zero otherwise), and an actual IPO indicator (equal to one if the company subsequently went public and zero otherwise). One might expect information related to companies in specific geographical locations or industries to likewise be more frequently reported than those which have been involved in M&A activities, or which went public at some point. The latter hypothesis is based on the assumption that once a company goes public, the database providers collect historic information and backfill it into their datasets. While the tables contain all VC databases, for the sake of brevity we focus our discussion below on findings on the highest-ranked sources: VS, PB and CB. All discussed relationships are at least statistically significant at the 5%.

Table 2.2.9: Determinants of a company appearing in the VC databases

Logistic regressions for the determinants of a company appearning in the VC databases for 108 companies which received financing between 1999 and 2019. The dependent variable equals 1 if the company is included in the respective database and equals 0 if the company is not covered. The independent variables are the natural logarithm of the actual total amount of capital raised in millions of Euros (lnTotalRaised), a count variable reflecting the founding year of the company of the founder (FoundingYear), an actual geography indicator (equal to 1 if the company is in Germany (DE), France (FR), United States (US) and 0 otherwise), an actual industry indicator (equal to 1 if software/IT (IT), life sciences, bio technology, healthcare and medical (LS) and 0 otherwise), a M&A indicator (equal to 1 if the company has been involved in M&A activities and 0 otherwise), and an actual IPO indicator (equal to 1 if the company subsequently went public and 0 otherwise). We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS). The values in parenthesis are t-test values, based on heteroskedasticity-robust standard errors. ***, **, and * indicate significance at 1%, 5% and 10%, respectively.

VARIABLES	AL	CB	CI	DR	PB	PQ	TR	VS
lnTotalRaised	0.248	0.495**	0.616**	0.519**	0.190	0.284	0.325	0.651**
	(0.215)	(0.239)	(0.254)	(0.231)	(0.231)	(0.185)	(0.217)	(0.289)
FoundingYear	0.162**	0.144***	0.281***	0.157***	0.195***	-0.0491	0.233***	-0.0249
	(0.0670)	(0.0476)	(0.0617)	(0.0474)	(0.0544)	(0.0418)	(0.0569)	(0.0560)
DE	-0.448	0.195	0.213	0.353	0.970	0.135	0.354	1.323**
	(0.597)	(0.560)	(0.603)	(0.549)	(0.626)	(0.470)	(0.578)	(0.636)
UK	4.334**	-	-	1.065	0.0115	1.337	-	-
	(1.864)	-	-	(1.533)	(1.564)	(1.333)	-	-
FR	-	-	-	-	-	0.680	-	-
	-	-	-	-	-	(1.388)	-	-
US	1.144	-0.0424	2.691**	1.214	0.249	-0.937	0.0868	-0.494
	(1.099)	(0.850)	(1.087)	(0.892)	(0.884)	(0.809)	(0.937)	(0.944)
ITSoftware	2.809*	1.587*	0.826	2.337***	0.836	1.186	1.904**	2.946**
	(1.510)	(0.863)	(0.906)	(0.858)	(0.852)	(0.766)	(0.967)	(1.168)
HealthBio	-0.178	1.109	1.506	1.967**	1.935*	-0.0959	2.898**	2.759**
	(1.722)	(0.938)	(1.098)	(0.937)	(1.040)	(0.824)	(1.164)	(1.242)
MA	1.305	1.277*	0.377	0.373	1.113	0.948	1.483**	2.079*
	(0.804)	(0.716)	(0.679)	(0.620)	(0.744)	(0.593)	(0.735)	(1.104)
IPO	-	0.182	0.244	1.729	-1.566	0.00787	-	0.0288
	-	(1.307)	(1.279)	(1.418)	(1.414)	(1.198)	-	(1.553)
Constant	-330.6**	-291.8***	-565.7***	-319.0***	-391.2***	96.98	-469.6***	46.63
	(135.1)	(95.73)	(124.3)	(95.52)	(109.4)	(83.97)	(114.7)	(112.6)
Observations	108	108	108	108	108	108	108	108

Company information. Table 2.2.9 describes the determinants of a company appearing in the VC databases. For VS, we find that the greater the total amount raised per company, the higher its likelihood to be included in the database. Moreover, IT/Software companies, Biotech/Medical/Healthcare companies and companies based in Germany are more likely to be included than others. VS does not exhibit any significant sampling biases related to founding years. PB, however, exhibits a significant time trend, as it is more likely to include younger companies than older ones. Our analysis does not reveal further significant biases for PB. However, CB reveals similarities to both of the previous databases, as it is more likely to include companies that have raised more capital and is also more likely to report younger companies than older ones. Contrary to our initial assumption and in line with Kaplan et al. (2002), there is no evidence that companies which went public exhibit any sampling biases.

Founders. Table 2.2.10 exhibits the determinants of a founder appearing in the VC databases. VS does not exhibit any significant biases, whereas PB is more likely to report founders whose companies raised more capital, were founded in recent years and have been involved in M&A activities. CB reveals a similarly positive time trend, and is more likely to include founders of companies that are classified as IT/Software. Again, the fact that companies went public does not seem to impact whether a founder is reported or not.

Funding. In Tables 2.2.11 to 2.2.13, we also include the natural logarithm of the actual financing round size (*lnRoundAmount*) and the natural logarithm of the actual financing round post-money valuation (*lnRoundPost*). Table 2.2.11 shows the determinants of a financing round appearing in the VC databases. All providers are consistently more likely to report greater financing rounds but are less likely to report those with higher post-money valuations, which is somewhat surprising. While PB and CB exhibit significantly positive time trends concerning the year of the financing round, VS shows

a negative time trend, indicating that older rounds are more likely to be included than younger ones. Similar to Table 2.2.10, PB is more likely to include financing rounds of companies that have been involved in M&A activities, and CB is more likely to include companies classified as IT/Software.

Table 2.2.10: Determinants of a founder appearing in the VC databases

Logistic regressions for the determinants of a founder appearning in the VC databases for 301 founders from 108 companies which received financing between 1999 and 2019. The dependent variable equals 1 if the founder is included in the respective database and equals 0 if the founder is not covered. The independent variables are the natural logarithm of the actual total amount of capital raised by the founder's company in millions of Euros (lnTotalRaised), a count variable reflecting the founding year of the company (FoundingYear), an actual geography indicator (equal to 1 if the company is in Germany (DE), France (FR), United States (US) and 0 otherwise), an actual industry indicator (equal to 1 if software/IT (IT), life sciences, bio technology, healthcare and medical (LS) and 0 otherwise), a M&A indicator (equal to 1 if the company has been involved in M&A activities and 0 otherwise), and an actual IPO indicator (equal to 1 if the company subsequently went public and 0 otherwise). We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS). The values in parenthesis are t-test values, based on heteroskedasticity-robust standard errors. ***, **, and * indicate significance at 1%, 5% and 10%, respectively.

VARIABLES	AL	CB	DR	PB	TR	VS
lnTotalRaised	0.225*	0.225*	0.350***	0.282**	0.300**	-0.0414
	(0.127)	(0.125)	(0.124)	(0.132)	(0.123)	(0.195)
FoundingYear	0.0810**	0.175***	0.141***	0.192***	0.172***	0.0286
	(0.0359)	(0.0314)	(0.0315)	(0.0335)	(0.0326)	(0.0468)
DE	0.801**	-0.462	-0.589*	0.571*	-0.0675	-0.509
	(0.364)	(0.344)	(0.335)	(0.344)	(0.331)	(0.503)
UK	3.041***	1.224	-1.330	1.053	0.948	-
	(1.163)	(1.235)	(1.262)	(1.204)	(1.111)	-
FR	-	-0.447	-0.396	0.00281	-1.265	-
	-	(1.166)	(0.943)	(1.160)	(0.894)	-
US	1.656**	0.439	-0.101	0.318	1.313**	-
	(0.730)	(0.580)	(0.659)	(0.578)	(0.635)	-
ITSoftware	2.868**	1.781***	3.205***	0.532	1.527***	1.539
	(1.164)	(0.547)	(0.844)	(0.527)	(0.562)	(1.146)
HealthBio	1.168	0.399	1.659*	-0.0321	-0.438	-
	(1.234)	(0.607)	(0.895)	(0.599)	(0.654)	-
MA	-0.437	0.809*	0.191	1.151***	0.527	1.058*
	(0.473)	(0.416)	(0.413)	(0.445)	(0.406)	(0.598)
IPO	-0.0766	0.987	1.835**	0.562	-	-
	(1.221)	(0.839)	(0.920)	(0.908)	-	-
Constant	-167.4**	-353.6***	-287.3***	-386.4***	-346.6***	-60.67
	(72.43)	(63.14)	(63.46)	(67.52)	(65.71)	(94.31)
Observations	301	301	301	301	301	301

Table 2.2.11: Determinants of a financing round appearing in the VC databases

Logistic regressions for the determinants of a financing round appearing in the VC databases for 339 financing rounds in 108 companies from July 1999 to June 2019. The dependent variable equals 1 if the round is included in the respective database and equals 0 if the round is not covered. The independent variables are the natural logarithm of the actual financing round size in millions of Euros (lnRoundAmount), the natural logarithm of the actual post-money valuation of the financing round in millions of Euros (lnRoundPost), the natural logarithm of the actual total amount of capital raised by a company in millions of Euros (lnTotalRaised), a count variable to reflect the year of the financing round (RoundYear), an actual geography indicator (equal to 1 if the company is in Germany (DE), France (FR), United States (US) and 0 otherwise), an actual industry indicator (equal to 1 if software/IT (IT), life sciences, bio technology, healthcare and medical (LS) and 0 otherwise), a M&A indicator (equal to 1 if the company has been involved in M&A activities and 0 otherwise), and an actual IPO indicator (equal to 1 if the company subsequently went public and 0 otherwise). We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS). The values in parenthesis are t-test values, based on heteroskedasticity-robust standard errors. ***, **, and * indicate significance at 1%, 5% and 10%, respectively.

VARIABLES	AL	CB	CI	DR	PB	PQ	TR	VS
lnRoundAmount	0.729***	0.729***	0.804***	0.681***	1.074***	0.768***	0.816***	0.830***
	(0.194)	(0.194)	(0.204)	(0.192)	(0.217)	(0.202)	(0.201)	(0.200)
lnRoundPost	-0.452**	-0.452**	-0.405**	-0.484**	-0.604***	-0.571***	-0.444**	-0.486**
	(0.192)	(0.192)	(0.197)	(0.191)	(0.208)	(0.201)	(0.195)	(0.201)
lnTotalRaised	0.139	0.139	0.0695	0.105	-0.0776	0.118	-0.0662	0.0366
	(0.140)	(0.140)	(0.141)	(0.137)	(0.145)	(0.136)	(0.137)	(0.141)
RoundYear	0.105***	0.105***	0.175***	0.122***	0.144***	-0.0646**	0.132***	-0.0575*
	(0.0259)	(0.0259)	(0.0301)	(0.0268)	(0.0283)	(0.0257)	(0.0279)	(0.0298)
DE	0.0739	0.0739	0.350	0.126	0.579*	0.135	0.298	0.553*
	(0.283)	(0.283)	(0.292)	(0.283)	(0.304)	(0.290)	(0.281)	(0.288)
UK	1.696	1.696	2.177*	0.300	-1.295	1.838*	2.050*	1.773
	(1.177)	(1.177)	(1.203)	(0.935)	(0.988)	(0.975)	(1.187)	(1.189)
FR	0.413	0.413	-	0.942	0.0548	0.158	-	-0.0750
	(1.167)	(1.167)	-	(1.162)	(1.183)	(0.924)	-	(0.983)
US	0.217	0.217	1.720***	1.269**	0.578	-0.324	-0.101	-0.168
	(0.474)	(0.474)	(0.529)	(0.496)	(0.512)	(0.487)	(0.488)	(0.506)
ITSoftware	0.845**	0.845**	0.715*	0.876**	0.350	1.005**	0.344	1.072**
	(0.382)	(0.382)	(0.406)	(0.386)	(0.403)	(0.402)	(0.386)	(0.425)
HealthBio	0.487	0.487	0.693	0.623	0.433	-0.745	0.222	1.054*
	(0.492)	(0.492)	(0.531)	(0.495)	(0.532)	(0.516)	(0.501)	(0.551)
MA	0.597*	0.597*	0.227	0.0754	0.901**	0.429	0.719**	0.782**
	(0.344)	(0.344)	(0.359)	(0.337)	(0.371)	(0.338)	(0.356)	(0.389)
IPO	-0.259	-0.259	0.0352	-0.133	-0.455	0.220	-0.706	-1.006*
	(0.570)	(0.570)	(0.593)	(0.571)	(0.591)	(0.564)	(0.653)	(0.573)
Constant	-210.7***	-210.7***	-353.9***	-246.4***	-288.2***	129.1**	-265.6***	115.2*
	(52.13)	(52.13)	(60.58)	(53.91)	(56.87)	(51.58)	(56.20)	(59.90)
Observations	339	339	339	339	339	339	339	339

Table 2.2.12 describes the determinants of a financing round size of the respective financing round of a company appearing in the VC databases. In line with Table 2.2.11, round sizes of larger financing rounds are more likely to be included than those of smaller ones across all three databases. PB and CB are less likely to report round sizes for financing rounds with greater post-money valuations than for those with lower ones. Again, PB and CB are more likely to include round sizes of younger rounds, whereas VS is more likely to report older ones. Moreover, PB is more likely to report round sizes of companies that have been involved in M&A activities. Most surprisingly, however, Table 2.2.12 reveals that financing round sizes of companies that went public are less likely to be reported across VS, PB and CB. For round sizes, this contradicts our previous backfill assumption.

Table 2.2.12: Determinants of a financing round size appearing in the VC databases

Logistic regressions for the determinants of a financing round size appearning in the VC databases for 339 financing rounds in 108 companies from July 1999 to June 2019. The dependent variable equals 1 if the round size is included in the respective database and equals 0 if the round size is not covered. The independent variables are the natural logarithm of the actual financing round size in millions of Euros (lnRoundAmount), the natural logarithm of the actual post-money valuation of the financing round in millions of Euros (lnRoundPost), the natural logarithm of the actual total amount of capital raised by a company in millions of Euros (lnTotalRaised), a count variable to reflect the year of the financing round (RoundYear), an actual geography indicator (equal to 1 if the company is in Germany (DE), France (FR), United States (US) and 0 otherwise), an actual industry indicator (equal to 1 if software/IT (IT), life sciences, bio technology, healthcare and medical (LS) and 0 otherwise), a M&A indicator (equal to 1 if the company has been involved in M&A activities and 0 otherwise), and an actual IPO indicator (equal to 1 if the company subsequently went public and 0 otherwise). We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS). The values in parenthesis are t-test values, based on heteroskedasticity-robust standard errors. ***, **, and * indicate significance at 1%, 5% and 10%, respectively.

VARIABLES	AL	CB	CI	DR	PB	PQ	TR	VS
lnRoundAmount	0.959***	0.959***	0.848***	1.026***	0.914***	0.986***	0.909***	0.795***
	(0.205)	(0.205)	(0.201)	(0.214)	(0.206)	(0.222)	(0.209)	(0.189)
lnRoundPost	-0.388**	-0.388**	-0.342*	-0.371*	-0.556***	-0.656***	-0.341*	-0.252
	(0.195)	(0.195)	(0.193)	(0.202)	(0.199)	(0.216)	(0.197)	(0.190)
lnTotalRaised	-0.0226	-0.0226	-0.0526	-0.0916	0.165	0.119	-0.0618	-0.0372
	(0.137)	(0.137)	(0.137)	(0.143)	(0.138)	(0.144)	(0.142)	(0.134)
RoundYear	0.0974***	0.0974***	0.131***	0.146***	0.103***	0.00458	0.130***	-0.0808***
	(0.0267)	(0.0267)	(0.0283)	(0.0305)	(0.0265)	(0.0262)	(0.0296)	(0.0274)
DE	-0.273	-0.273	0.00257	-0.136	0.189	-0.145	0.0933	0.243
	(0.283)	(0.283)	(0.284)	(0.294)	(0.284)	(0.308)	(0.290)	(0.281)
UK	1.608	1.608	2.180*	0.180	-0.482	2.122**	2.256*	1.813
	(1.180)	(1.180)	(1.198)	(0.935)	(0.967)	(0.976)	(1.185)	(1.166)
FR	0.818	0.818	0.243	1.320	0.0327	0.0749	-0.620	0.558
	(1.173)	(1.173)	(0.955)	(1.188)	(0.977)	(0.951)	(0.920)	(0.963)
US	-0.178	-0.178	0.976**	1.297**	0.929*	-0.0374	-0.00430	-0.383
	(0.478)	(0.478)	(0.489)	(0.514)	(0.503)	(0.502)	(0.500)	(0.486)
ITSoftware	0.491	0.491	0.468	0.402	0.521	0.880**	0.301	0.942**
	(0.387)	(0.387)	(0.401)	(0.414)	(0.394)	(0.421)	(0.401)	(0.391)
HealthBio	-0.00738	-0.00738	0.907*	-0.0800	0.953*	-0.648	-0.127	0.566
	(0.504)	(0.504)	(0.518)	(0.528)	(0.531)	(0.551)	(0.530)	(0.496)
MA	0.378	0.378	0.417	-0.0643	0.855**	0.861**	0.864**	0.198
	(0.345)	(0.345)	(0.353)	(0.367)	(0.349)	(0.357)	(0.365)	(0.346)
IPO	-1.469**	-1.469**	-0.286	-1.027	-2.031**	1.052*	-2.105*	-1.230**
	(0.725)	(0.725)	(0.623)	(0.733)	(0.819)	(0.587)	(1.089)	(0.581)
Constant	-196.4***	-196.4***	-264.4***	-294.8***	-208.8***	-10.58	-262.7***	161.7***
	(53.79)	(53.79)	(56.97)	(61.36)	(53.38)	(52.61)	(59.56)	(55.16)
Observations	339	339	339	339	339	339	339	339

Table 2.2.13 exhibits the determinants of a post-money valuation appearing in the VC databases. These findings might be relevant for studies calculating returns based on post-money valuations. CB is not included in this analysis as it does not report post-money valuations. In line with aforementioned findings, we find that post-money valuations are significantly more likely to be reported for larger financing rounds. Although one might assume that round sizes correlate with post-money valuations, we find no significant bias with respect to the size of the post-money valuation itself. Concerning the time trend, we would again testify that PB is more likely to report post-money valuations of companies founded in more recent years, whereas VS is more likely to report those from earlier years in our sample. While VS is less likely to include post-money valuations of US based companies, PB is significantly more likely to report them for companies based in the US than for other countries. Similar to financing round

size coverage, VS is less likely to report post-money valuations for companies that went public than for private companies, which again contradicts the backfill hypothesis.

Table 2.2.13: Determinants of a post-money valuation appearing in the VC databases

Logistic regressions for the determinants of a post-money valuation appearning in the VC databases for 339 financing rounds in 108 companies from July 1999 to June 2019. Databases which do not provide financing rounds are omitted from the table. The independent variables are the natural logarithm of the actual financing round size in millions of Euros (lnRoundAmount), the natural logarithm of the actual post-money valuation of the financing round in millions of Euros (lnRoundPost), the natural logarithm of the actual total amount of capital raised by a company in millions of Euros (lnTotalRaised), a count variable to reflect the year of the financing round (RoundYear), an actual geography indicator (equal to 1 if the company is in Germany (DE), France (FR), United States (US) and 0 otherwise), an actual industry indicator (equal to 1 if software/IT (IT), life sciences, bio technology, healthcare and medical (LS) and 0 otherwise), a M&A indicator (equal to 1 if the company has been involved in M&A activities and 0 otherwise), and an actual IPO indicator (equal to 1 if the company subsequently went public and 0 otherwise). We present results for Angellist (AL), CBInsights (CI), Crunchbase (CB), Dealroom (DR), Pitchbook (PB), Preqin (PQ), Tracxn (TR) and VentureSource (VS). The values in parenthesis are t-test values, based on heteroskedasticity-robust standard errors. ***, **, and * indicate significance at 1%, 5% and 10%, respectively.

VARIABLES	CI	DR	PB	VS
lnRoundAmount	0.664**	1.272***	0.686**	0.784***
	(0.287)	(0.318)	(0.278)	(0.190)
lnRoundPost	-0.494*	0.117	-0.179	-0.274
	(0.278)	(0.279)	(0.271)	(0.191)
lnTotalRaised	0.457**	-1.192***	-0.202	0.0519
	(0.192)	(0.258)	(0.205)	(0.135)
RoundYear	0.0610	0.140***	0.134***	-0.0679***
	(0.0413)	(0.0446)	(0.0475)	(0.0261)
DE	-0.430	-0.167	-0.251	0.239
	(0.447)	(0.389)	(0.404)	(0.282)
UK	1.185	1.637	0.217	0.971
	(1.224)	(1.121)	(1.233)	(0.943)
FR	-	0.975	0.696	0.937
	-	(1.039)	(1.021)	(0.966)
US	0.678	0.939	1.930***	-1.136**
	(0.642)	(0.662)	(0.663)	(0.500)
ITSoftware	1.123*	0.0747	0.539	0.714*
	(0.631)	(0.641)	(0.673)	(0.379)
HealthBio	-0.323	-0.327	-0.185	0.542
	(0.787)	(0.762)	(0.787)	(0.476)
MA	0.790	0.0257	-0.486	0.0668
	(0.548)	(0.513)	(0.565)	(0.336)
IPO	0.254	-0.184	-	-1.322**
	(0.888)	(1.151)	-	(0.603)
Constant	-126.6	-282.0***	-271.5***	135.4***
	(83.11)	(89.72)	(95.54)	(52.52)
Observations	339	339	339	339

Overall, in our view the results presented in Tables 2.2.9 to 2.2.13 have several relevant implications: The fact that PB and CB exhibit positive time trends concerning the coverage of all funding related information, whereas the coverage in VS decreases over time, might to some degree explain why VS became academically less relevant after 2012 and why the other two databases have taken over the leading position as of today, as described in Table 2.2.1. Considering that previous database benchmarking studies focused solely on funding-related information, and that the majority of academic

studies based on the VC databases can be classified as finance-related research, the negative funding related time trends of VS and the increasing coverage of PB and CB seem to become even more plausible as an explanation. Furthermore, we find that greater financing rounds are more likely to be reported than lower ones. Similarly, financing round sizes and post-money valuations are more likely to be reported for greater financing rounds than for lower ones. Although, all providers over- or underreport companies with database-specific characteristics such as geography or industry classifications, there seem to be no other consistent patterns or biases across databases. Neither the total amount raised, nor an IPO or M&A event, consistently impact whether a company, founder, financing round, round size or post-money valuation is reported.

2.2.5 Summary and implications

We compare the actual contracts and investment documentation of 339 VC financing rounds from 396 investors in 108 different companies with their characterisation in the eight most relevant VC databases. Our results should help academic scholars and VC practitioners to better understand the coverage and quality of their datasets, and thus interpret the results more accurately. Specifically, with respect to the increasing efforts of leveraging machine learning in VC, our work should help to increase the representativeness of the training data and remove potential biases from the models.

Our analysis reveals that VS, PB and CB have the best coverage, and are the most accurate databases across the dimensions of general company, founders and funding information. A combined dataset with the best possible coverage would consist of general company information from VS, founder information from PB and funding information from a combination of CB, PB and VS.

Concerning sampling biases, we find that greater financing rounds are more likely to be reported than lower ones. Similarly, financing round sizes and post-money valuations are more likely to be reported for greater financing rounds than for lower ones. Although our results reveal a variety of further sampling errors and biases, we cannot summarize them into consistent patterns across all databases. In any case, our analysis underlines that scholars need to be cautious in picking data sources for a given project, as this choice might materially impact research results. As a side note, CB and PB seem to reveal very similar biases, which indicates that these providers might potentially rely on the same or similar raw data sources.

We are well aware that, like all research efforts, our work is conditioned by several limitations. To that effect, we have consistently tried to ensure the robustness and validity of our research with respect to the following two issues. Firstly, our original dataset consists of 339 VC financing rounds in 108 companies only. Certainly, it is not representative for all companies and financing rounds across Europe, and certainly not globally, but it serves as a suitable approximation to better understand the coverage, quality and biases of the VC databases for companies with similar profiles. Secondly, we have tried to extend previous research by adding further dimensions such as general company and founder-related information, but are aware that there are even more variables to be considered and to be challenged. Though these variables are not covered in the original documents provided to us, we suggest that future research finds ways to collect such original data and benchmarks the databases against it.

2.3 Essay 3 – Human Versus Computer: Benchmarking Venture Capitalists and Machine Learning Algorithms for Investment Screening

Abstract

I conduct an investment screening performance benchmarking between 111 venture capital (VC) investment professionals and a supervised gradient boosted tree (or "XGBoost") classification algorithm to create trust in machine learning (ML)-based screening approaches, accelerate the adoption thereof and ultimately enable the traditional VC model to scale. Using a comprehensive dataset of 77,279 European early-stage companies, I train a variety of ML algorithms to predict the success/failure outcome in a 3- to 5-year simulation window. XGBoost algorithms show particularly excellent performance in terms of accuracy and recall, which denote the most important metrics in my setup. I benchmark the performance of the selected algorithm against that of the VC investment professionals by providing equal information in the form of 10 company one-pagers via an online survey and requesting respondents to select the five most promising companies for further evaluation. In addition to finding characteristic-specific performance dependencies for VCs, I find that the XGBoost algorithm outperforms the median VC by 25% and the average VC by 29%. Although I do not suggest replacing humans with ML-based approaches, I recommend an augmented solution where intelligent algorithms narrow down the upper part of the deal-flow funnel, allowing VC investment professionals to focus their manual efforts on the lower part of the funnel. Using this approach, they can rely on a scalable but objective pre-selection and focus their manual resources on evaluating the most promising opportunities and putting themselves into the best position to secure these deals.

Keywords: Venture capital, machine learning, performance, benchmarking

Authors: Andre Retterath

First Author: Andre Retterath

Current Status: Working paper

2.3.1 Introduction

For more than a decade, near-zero interest rates have prompted investors to attempt to identify alternative asset classes with attractive return profiles. As a consequence, the venture capital (VC) industry, which serves as a form of private capital management, has seen a significant jump in capital inflow, resulting in an increased number of funds being raised and fund sizes becoming considerably larger. For example, the sum of United States (US)-based and European VC funds raised per year increased by 33% from 251 in 2009 to 336 in 2019. Simultaneously, the median fund size increased by 57% from $50 million in 2009 to $78.5 million in 2019 in the US and by 425% from €20 million in 2009 to €105 million in 2019 in Europe (Pitchbook, 2019; Pitchbook and NVCA, 2019). The almost constant rate at which companies are being founded globally and the steady number of deals being made across all stages (Lavender, Moore, Smith, and Eli, 2019), however, create an asymmetry between the amount of capital to be deployed and the number of investment opportunities available. Inevitably, the competition amongst VCs to identify the most promising investment opportunities as early as possible and to put themselves into a position to secure deals against other investors has gradually intensified.

Alongside manual deal origination practices such as active outbound outreach via investment professionals or passive inbound deal flow[20] attracted by an investor's brand, VCs have started to leverage a variety of "*quantitative sourcing tools*" (Paul A Gompers et al., 2020). For example, web crawlers or external data feeds enhance their outbound deal flow and maximize their coverage. Although still early in the adoption of data-driven sourcing approaches, VCs increasingly utilize these tools because they complement their existing sourcing efforts and increase overall deal flow without known disadvantages or risks (Schmidt, 2019). Anecdotal evidence shows that at this stage of the sourcing process, it is mainly about

[20] "Deal flow" refers to the steady flow of (inbound and outbound) identified investment opportunities or deals. The terms "deals" and "investment opportunities" have the same meaning and are used interchangeably.

the quantity, rather than the quality, of additional opportunities identified. Considering that the manual screening process was already characterized by high amounts of available information (Zacharakis and Meyer, 1998) and intense time pressure (Zacharakis and Shepherd, 2001) two decades ago, the surge in automatically identified deal flow has exacerbated the impact of these factors and revealed the resource limitations of this traditional approach. When using manual screening processes, VCs are incapable of handling the increasing amount of available information, which forces them to either skip and not evaluate a growing number of potential targets or to rush through the data and thereby increase the risk of misclassification. Moreover, the existing literature shows that even in the absence of time pressure, VCs suffer from availability bias, similarity bias and overconfidence bias, which lead them to make subjective and thus suboptimal screening decisions (Franke, Gruber, Harhoff, and Henkel, 2006; Zacharakis and Meyer, 1998; Zacharakis and Shepherd, 2001). As a result, VCs face a high risk of overlooking promising opportunities.

To overcome the lack of scalability and reduce the potential for biased decisions during the screening stage, VCs and academics have started to explore several efficiency boosters and objectivization approaches ranging from deterministic scorecards to artificial intelligence (AI) or ML[21]-based screening tools. However, in contrast to quantitative sourcing approaches, these automated screening tools face significant skepticism because they actively narrow the deal funnel and thus have a direct impact on ultimate fund performance. Unlike quantitative sourcing, there is a significant risk associated with the use of automated screening in terms of unobserved misclassification.

As has been widely recognized, VC is an outlier business and mistakenly passing on the next Google or Facebook could make the difference between a firm falling within the top decile of VC companies and the rest (Paul A Gompers and Lerner, 1999; Kaplan and Schoar, 2005;

[21] "AI" can be defined as using a computer to mimic human behavior in some way, whereas "ML" is a subset of AI which consists of techniques that enable computers to become more accurate at predicting outcomes without being explicitly programmed to do so.

Korteweg and Sorensen, 2010; Retterath and Kavadias, 2020). VCs can lose their initial investment only once should a company go bust (i.e., a full write-off) but can multiply their invested capital more than once should a company take off. This asymmetric return profile leads them to not accept false negatives (FNs), which refers to classifying an investment as unsuccessful despite the fact that it would have become a success. Passing on desirable deals is generally heavily punished, and, as a consequence, VCs are reluctant to surrender control of their selection process and thus refrain from adopting automated screening tools. I assume that this "*automation–control trade-off*" prevents the traditional VC investment process from scaling and thus conducted 63 informal interviews throughout 2018 and 2019 to better understand the root cause of VCs' hesitation. In line with Schmidt's (2019) conclusion that "*In order to implement AI [in the investment process], organizational behaviors have to be changed slowly. The more often investment professionals are outperformed by the algorithm, the more the trust in the algorithm increases,"* I found that 95% of VCs refrain from adopting such tools because they are reluctant to sacrifice screening performance compared to their status quo. Therefore, I seek to resolve the automation–control trade-off by creating the necessary trust through a comprehensive white-box benchmarking study.

Following the recent literature, I train a variety of ML algorithms with a comprehensive dataset of 77,279 European early-stage companies to predict the success/failure outcome in a 3- to 5-year simulation window. XGBoost algorithms show particularly excellent performance in terms of accuracy and recall, which are the most important metrics in my setup. I benchmark the selected algorithm with VC investment professionals by providing equal information in the form of 10 company one-pagers via an online survey and requesting that participants select the five most promising companies for further evaluation. My results show that the XGBoost algorithm outperforms the median VC by 25% and the average VC by 29%. Although I do not suggest replacing humans with ML-based approaches, I recommend an augmented solution where intelligent algorithms narrow the upper part of the funnel from an unmanageable to a

manageable number of investment opportunities, thus allowing VC investment professionals to focus their manual efforts on the lower part of the funnel. Using this approach, they can rely on a scalable but objective pre-screening technique and focus their manual resources on further evaluating the most promising opportunities and putting themselves into the best position to win these deals. Moreover, I find characteristic-specific performance dependencies for subgroups of VCs. For example, I find that institutional VCs perform 33% better than the *median* corporate VC (CVC) and 22% better than the *average* CVC. In addition, I find a negative correlation between VC experience and screening performance after approximately a decade of VC experience, which might be explained by reliance on patterns as well as confirmation and availability biases. My results further show that the level of a venture capitalist's highest academic degree is positively correlated with screening performance and that science, technology, engineering and math (STEM) graduates perform better than business graduates and graduates with other degrees.

The remainder of this paper is structured as follows: Section 2.3.2 describes the traditional VC investment process and its limited ability to cope with the ever-increasing challenges associated with objectively selecting the most promising investment opportunities. I further explain how automated screening approaches might resolve the bottlenecks of the traditional VC model but have not been widely adopted due to the so-called "*automation–control trade-off*." As investors simply do not trust a black box algorithm when it comes to the most critical aspect of their business, I assume that, in particular, the lack of suitable performance benchmarking studies prevents large-scale adoption of such tools. Section 2.3.3 explores a variety of performance benchmarking approaches that help to shed light on how these tools perform in comparison with the manual selection by VC investment professionals (i.e., the status quo). I conclude this chapter by detailing my novel benchmarking approach, the test sample and the participating VC investment professionals. Section 2.3.4 complements the preceding chapters by explaining how I train and select the best-performing ML algorithm for

the purpose of VC investment screening. Conflating both streams, Section 2.3.5 presents the results of my investment screening performance benchmarking between VC investment professionals and the XGBoost classifier. I discuss the results in detail in Section 2.3.6 before identifying the limitations of this study and potential directions for future research.

2.3.2 Venture capital investment process and automation approaches

To set the stage and provide context for the subsequent discussion, I start this chapter by describing the VC investment process and related trends. I also identify the processual bottlenecks that prevent the traditional VC process from being scaled and explain approaches to removing them, which mainly rely on objectivization and automation.

2.3.2.1 Investment process

The VC investment process can be described as a multistage sequential process. Table 2.3.1 summarizes the models established by Boocock and Woods (1997), Fried and Hisrich (1994), Hall and Hofer (1993), Tyebjee and Bruno (1984) and Wells (1974), respectively. I follow Fried and Hisrich's (1994) six-stage model, which consists of (1) "*deal origination*," (2) "*firm-specific screen*," (3) "*generic screen*," (4) "*first-phase evaluation*," (5) "*second-phase evaluation*" and (6) "*closing*" to describe the upper part (1) to (3) of the process, which is considered to drive approximately 60% of VC returns (Sorenson, 2007) and is the main field of application for automation and AI/ML in VC (Schmidt, 2019), in detail below.

The (1) "*deal origination"* stage refers to becoming aware of new investment opportunities or filling up the deal funnel. A wide funnel is a precondition for making appropriate investments and achieving outsized returns (Sahlman, 1990). According to Gompers, Gornall, Kaplan, and Strebulaev (2020), 58% of investment opportunities are referred by a venture capitalist's network, 30% of which come from professional networks, 20% from other investors and 8% from their existing portfolio companies.

Table 2.3.1: Stages of the VC investment process

Overview of the five most established VC decision making process models. Taken and modified from Hall/Hofer (1993) and Boocock/Woods (1997).

Stages	Wells (1974)	Tyebjee/ Bruno (1984)	Hall (1989)	Fried/ Hisrich (1994)	Boocock/ Woods (1997)
1.	Search	Deal origination	Generting deal-flow	Deal origination	Generting deal-flow
2.	Screen	Screening	Proposal screening	Firm-specific scree	Initial screening
3.	n/a	n/a	Proposal assessment	Generic screen	First meeting
4.	Evaluation	Evaluation	Project evaluation	First phase evaluation	Second meeting
5.	n/a	n/a	n/a	n/a	Board presentation
6.	n/a	n/a	Due diligence	Second phase evaluation	Due diligence
7.	n/a	Deal structuring	Deal structuring	Closing	Deal structuring
8.	Operations	Post-investment activities	Venture operations	n/a	Monitoring of investments
9.	Cashing out	n/a	Cashing out	n/a	Cashing out

The authors find that while only 10% of companies come inbound via cold emails or website submission forms, approximately 32% of a venture capitalist's deal flow is self-generated. Through the rise of digitization and the increasing availability of free information, VCs have started to collect online information from startup databases such as Crunchbase or Pitchbook and to proactively approach suitable targets. Unfortunately, these manual efforts do not scale, as there is an approximately proportional relationship between the time spent engaging in desk research and the number of opportunities identified. VCs simply cannot allocate the resources required to manually search for and identify all promising opportunities. To address this issue, they have started to leverage automated web crawlers that continuously seek to identify new investment opportunities through a variety of sources, such as public registers, news mentions, product platforms, app stores and many more. As these "*quantitative sourcing*" (Paul A Gompers et al., 2020) approaches lead to a significant increase in deal flow and do not seem to be associated with any kind of disadvantages or *direct* risks, the adoption of such solutions, both external and home-made, is continuously rising. However, publicly

sourced and unverified information bears the *indirect* risk of serving as the basis for poor decisions in the subsequent screening or evaluation process (Kaplan and Lerner, 2016; Kaplan et al., 2002).

The (2) "*firm-specific screen"* stage focuses on reducing the number of investment opportunities or narrowing the deal funnel by applying so-called "hard" selection criteria (SC[22]) such as geography, stage or industry. These hard SC frequently derive from the investment strategy that VCs and their limited partners (LPs) have mutually agreed upon and provide little scope for interpretation. Therefore, this step requires limited intellectual capacity and could, if necessary, easily be automated through a deterministic decision tree.

The purpose of stage (3) "*generic screen*" is to separate the wheat from the chaff and select the most promising opportunities, which are thoroughly assessed in the subsequent evaluation phase. VCs narrow the funnel through the application of a range of generic or "soft" SC that are intended to help them separate potentially successful investments from the rest. The ultimate go/no-go decision for further evaluation is based on two dimensions: a) the perceived strength and weakness profile within one specific SC, which is referred to as the *micro level*, and b) the relative weighting (important versus unimportant) of the various SC within the overall go/no-go decision, which is referred to as the *macro level*. In contrast to the (2) "*firm-specific screen*," the (3) "*generic screen*" provides significant scope for interpretation and subjectivity (Paul A Gompers et al., 2020). Knowing that VCs suffer from availability bias (Zacharakis and Meyer, 1998), overconfidence bias (Zacharakis and Shepherd, 2001) and similarity bias (Franke et al., 2006), among others, and assuming that less than 10–20% of investment opportunities make it past the screening phase (Dixon, 1991; Petty and Gruber, 2011), I consider the (3) "*generic screen*" as the most critical but also most vulnerable stage of the VC decision-making process. To better understand the status quo, the bottlenecks and the

[22] Note that SC means selection criteria but also represents "features" or independent variables in the ML context. I use the abbreviation interchangeably.

automation potential in the (3) "*generic screen*" stage, I describe its most important components in more detail below.

2.3.2.2 Selection criteria

Table 2.3.2 summarizes the extensive literature regarding the predictive power of different SC used in the VC selection process. Prior studies have applied a variety of methods and distinguished the SC into different number of categories by using different definitions and variables. At the most basic level, however, the SC can be classified as falling into five main categories:

(1) General company: This category includes mostly hard SC, such as headquarters location and financing stage or industrial focus of a company, and allows the investor to obtain a first impression of what a company does, its history and whether there is a fit with the venture capitalist's portfolio. Although this category has not been addressed in many studies, scholars agree on its relative importance for the overall decision-making process (Hall and Hofer, 1993).

(2) Market: This problem-oriented category includes SC such as market size, market growth, market potential, character of the market (niche, fragmented, etc.) and character of the economic environment (rapid, chaotic, etc.) to enable a venture capitalist to better understand the actual customer pain points and the potential of a business. As VCs educate themselves about specific markets and develop expertise over time, market-related SC information provided by companies becomes less important for the overall screening process (MacMillan, Siegel, and Narasimha, 1985; Muzyka, Birley, and Leleux, 1996; Stuart and Abetti, 1987; Wells, 1974).

(3) Product/service: This solution-oriented category includes important soft SC such as product attributes, product strategy, patentability, uniqueness, differentiation, technical edge of a product or stage of product development, all of which help a venture capitalist

to better understand what a company is offering and how innovative its product or service is. This category is individual for every company and is considered to be one of the most important generic SC (Bachher and Guild, 1996; Hisrich and Jankowicz, 1990; Siskos and Zopounidis, 1987; Jeffrey A Timmons, Muzyka, Stevenson, and Bygrave, 1987; Tyebjee and Bruno, 1984).

(4) Entrepreneurial team: This category summarizes founder- or management-related SC, such as management skills, leadership capabilities, business qualifications, technical qualifications, education, gender and age, and is intended to help a venture capitalist develop a sense of the people behind a business. Scholars and practitioners agree that particularly in early-stage ventures, the founding team is the most important success factor (Franke et al., 2006; Paul A Gompers et al., 2020; MacMillan et al., 1985; MacMillan, Zemann, and Subbanarasimha, 1987; Rah, Jung, and Lee, 1994; Wells, 1974).

(5) Funding: This category summarizes funding- and shareholder-related SC, such as existing investors, valuation, deal structure or use of funding proceeds, and helps a venture capitalist to evaluate the financial attractiveness of a particular investment opportunity (Muzyka et al., 1996; Petty and Gruber, 2011; Poindexter, 1976). With respect to the signaling effects of existing investors, this SC can quickly lead an inspecting venture capitalist to make a go/no-go decision concerning further evaluation. For example, should a top-tier venture capitalist have already invested in a company's seed round, an inspecting venture capitalist is likely to pursue the opportunity because the involvement of a top-tier investor is often associated with success. Consequently, this group of criteria can be considered rather important for the screening process.

Table 2.3.2-1: VC investment selection criteria

Overview of the literature on VC selection criteria across the five major groups of "general company", "market", "product/service", "entrepreneur/team" and "funding" information. "X" in a cell means that the author in the respective column has researched or discuss the selection criteria in the respective line. Asterisks (*) are taken from Martel (2006). Others are original content.

Study	*Wells (1974)	*Poindexter (1976)	*Ruby (1984)	*Tyebjee/Bruno (1984)	*Siskos/Zopounidis (1987)	*Timmons et al. (1987)	*Rah et al. (1994)	*Boehm (2002)	MacMillan/Siegel/Narasimha (1985)	Stuart/Abetti (1987)	MacMillan/Zemann/Subbanarasimha (1987)	Robinson (1987)	Rea (1989)	Hisrich/Jankowicz (1990)	Roure/Keeley (1990)	Dixon (1991)	Hall/Hofer (1993)	Fried/Hisrich (1994)	Muzyka et al. (1996)	Bachher/Guild (1996)	Boocock/Woods (1997)	Zacharakis/Meyer (2000)	Beim (2004)	Kaplan/Strömberg (2004)	Martel (2006)	Henkel et al. (2008)	Petty/Gruber (2009)	Kollmann/Kuckertz (2009)	Gornall et al. (2016)	Gompers et al. (2020)
General Company																														
Geographic Location										X							X		X		X									
VC Focus/Fit (product, stage, size…)									X		X						X		X								X			X
VC Portfolio/Fit investment strategy																											X	X		X
VC External Info Source																	X										X			
Continuity of Company (History)														X																
Comparable Companies (Description)																													X	
Completeness of Information (full, lack…)															X		X			X										
Content (easy to read, length…)																	X			X										
Market	X															X			X	X		X		X				X	X
Market Size	X		X	X		X	X	X		X							X	X	X									X		
Market Growth/Potential	X		X	X		X	X		X	X	X	X	X			X	X	X	X		X				X		X	X		
Character or Market (niche, fragment…)																											X			
Economic Environment (rapid, chaotic...)										X							X			X										
Regulations																											X			
Projected Market Share										X					X				X											
Competitive Strength/Number						X		X	X		X				X		X								X					
Competition										X		X							X	X										X
Sensitivity to Business Cycles				X	X														X											
Seasonality of Market-Product																			X											
Buyer Concentration															X															
Venture Creates New Market									X		X								X											
Product/Service	X																					X							X	X
Product Attributes	X			X			X	X	X	X	X							X		X										
Strategy/Model										X							X										X			
Proprietarity/Patentability			X	X		X	X		X	X	X	X								X							X	X		
Uniqueness/Differentiation	X		X	X	X	X	X			X				X					X	X	X						X	X		
Technical Edge/Innovation	X			X			X							X											X			X		
Stage of Development		X		X	X										X						X				X					
Technology Life Cycle				X		X				X																				
Expected Profit Margin				X						X																				
Project Growth in Turnover																X														
Resistance to Risk				X																					X					
Scalability																									X					
Barriers/Ease to Entry				X		X				X									X						X					
Product Superiority							X						X		X					X				X						
Existing Customer Base																			X						X					
Market Acceptance/Interest	X			X			X		X	X	X									X					X		X	X		
Potential for Partnerships																			X						X					
Prototype/R&D Level					X	X			X	X	X									X							X			
Value for Customers (cost:benefit…)										X																				

Table 2.3.2-2: VC investment selection criteria

Overview of the literature on VC selection criteria across the five major groups of "general company", "market", "product/service", "entrepreneur/team" and "funding" information. "X" in a cell means that the author in the respective column has researched or discuss the selection criteria in the respective line. Asterisks (*) are taken from Martel (2006). Others are original content.

Study	*Wells (1974)	*Poindexter (1976)	*Ruby (1984)	*Tyebjee/Bruno (1984)	*Siskos/Zopounidis (1987)	*Timmons et al. (1987)	*Rah et al. (1994)	*Boehm (2002)	MacMillan/Siegel/Narasimha (1985)	Stuart/Abetti (1987)	MacMillan/Zemann/Subbanarasimha (1987)	Robinson (1987)	Rea (1989)	Hisrich/Jankowicz (1990)	Roure/Keeley (1990)	Dixon (1991)	Hall/Hofer (1993)	Fried/Hisrich (1994)	Muzyka et al. (1996)	Bachher/Guild (1996)	Boocock/Woods (1997)	Zacharakis/Meyer (2000)	Beim (2004)	Kaplan/Strömberg (2004)	Martel (2006)	Henkel et al. (2008)	Petty/Gruber (2009)	Kollmann/Kuckertz (2009)	Gornall et al. (2016)	Gornall et al. (2016)
Entrepreneur/Team					X	X			X		X	X	X			X	X	X								X			X	X
Management Skills/Leadership	X	X	X	X	X	X	X		X	X	X	X		X	X	X	X		X	X	X						X	X	X	X	X
Business Qualification																			X	X								X		
Technical Qualification												X							X	X								X		
Education																										X				X
Age																										X				
Completeness of Team									X	X		X			X		X		X	X						X	X			X
Marketing Skills							X								X	X			X											
Management Financial Skill		X														X			X											
Management Stake in Firm				X			X									X														
Articulate About Venture									X	X	X	X																		
Willingness to Cooporate with VC																	X			X										
Personal Motivation	X						X					X																		
Effectiveness										X																				
Confidence																											X			
Capable of Sustained Effort/Commitment	X						X		X		X									X								X		
Ability to Evaluate Risk							X		X		X									X										
Relevant Track Record				X		X	X		X	X	X	X		X	X				X		X								X		X
Market Familiarity				X		X	X		X	X	X									X										X
Entrepreneur Personality							X		X	X	X	X						X		X							X	X	X	X
Key-man Issues																			X								X			
References	X											X						X												
Reputation									X		X																X			
Funding																			X			X		X	X			X	X
Cash-out Method	X			X	X												X		X											
Expected Rate of Return		X		X		X			X		X		X				X	X	X	X							X	X		X
Expected Risk		X											X							X										X
Time to Payback																			X											
Time to Breakeven																			X											
Percentage of Equity		X											X							X										X
Investor Provisions		X																		X										
Openness to Investment Influence																			X											
Size of Investment	X			X												X		X	X		X									X
Use of Proceeds																											X			
Funding Base														X																
Liquidity of Investment/Exit	X			X					X		X	X													X		X	X	X	
Valuation																									X		X		X	X
Deal Structure (lack of investor…)																			X								X			X
Realism (financial etc.)																	X			X										

As different studies research different SC and employ different methods, studies may vary in how they determine the importance of macro SC with regard to the overall go/no-go decision and/or weighted effect on venture success (Paul A Gompers et al., 2020). It becomes clear that the micro-level perception of different SC, as well as the respective macro-level weighting, is subject to individual perspectives. The combination of subjectivity across both SC levels and multiple investment team members being involved in the selection process leads to varying screening outcomes, particularly in early-stage investing, where data is typically more qualitative.

To cope with this ambiguity, early-stage VCs tend to apply an internally agreed-upon set of heuristics that is based on previous experiences and is intended to objectivize the decision-making process. While this approach certainly works for hard criteria and quantitative metrics, it becomes more difficult to apply when evaluating soft criteria such as team setup, degree of innovation or competitive landscape. Most VCs are aware of potential misclassification and thus discuss the most interesting opportunities in weekly deal-flow calls. Still, more than 55% of go/no-go screening decisions are made by only one investment team member (Franke et al., 2006). In addition to subjectivity and the significant risk of misclassification, the recent rise of automation in the deal origination phase has pushed the traditional screening process to its limits. Investors can simply no longer handle the number of identified investment opportunities manually. As a consequence of inconsistent decisions and limited scalability, practitioners and academics have started to explore innovative approaches to separating the weed from the chaff. The most important once are described below.

2.3.2.3 Structured screening approaches

Manual scorecards. The purpose of scorecards is to leverage a fixed set of rules that allow different investors to quantify SC and thereby increase the likelihood of obtaining the same result. On the micro level, a scorecard for every individual SC that contains a specific definition

for each score exists; for example, a 0 (= worst) to 3 (= best) scorecard for a team would define 0 points as indicating no academic background and no relevant experience, 1 point as undergrad background or up to 2 years' experience, 2 points as graduate background or 3–5 years of relevant experience and 3 points as graduate background and more than 5 years' experience. A similar model is that established by Kirsch, Goldfarb, and Gera (2009), which features 22 variables and is used to predict fundraising success. Alternatively, there exist binary scores for every SC (e.g., "good" versus "bad" team). Although these models ultimately lead to less differentiated outcomes, they have been widely adopted due to their simplicity. For example, Lussier (1995) developed a widely used 15-variable model for predicting the failure or success of non-financial business ventures that has been applied across industries and geographies. For most practitioners, the definitions and characteristics used for computing scores are based on heuristics that they have gained through past experiences, which are occasionally complemented through insights from the literature. Once a scorecard is defined, the outcome for every SC should be consistent for different investment team members within a specific firm. Subsequently, the individual SC scores are manually aggregated on the macro level and weighted based on a fixed formula (e.g., team has a weight of 1.0, market has a weight of 1.5, business model has a weight of 0.5, traction has a weight of 1.0, etc.). Alternatively, there exist unweighted scores that sum individual SC without weightings, which, again, leads to less differentiated outcomes. The resulting company scores help investors to rank multiple opportunities and to focus their resources on the most promising ones. Manual scorecards enjoy increasing adoption across inbound and outbound deal flow, as they reduce subjectivity and can be applied to any kind of data format while also making it possible to increase consistency and efficiency without surrendering control over the process.

Automated scorecards. While, for manual scorecards, investors need to manually compute every SC score and then manually aggregate them into a combined company score, automated scorecards leverage the same set of static rules but are fully automated across the micro and

macro levels. On the micro level, the approach relies on a variety of natural language understanding (NLU) models that identify specific keywords in the data and match them with a pre-defined weighted dictionary; for example, if a model identifies a tier 1 university in the founder's curriculum vitae, a maximum score of 3 is assigned, whereas, for tier 2, tier 3 and other universities, a value of 2, 1 or 0 is assigned, respectively. The process flow is the same as for manual scorecards, but the input data processing is automated, which yields significant efficiency gains but requires structured data. Although it would theoretically be possible to transform unstructured inbound deal-flow information in the form of emails and pitch decks into structured tables, practice shows that doing so is highly complex and, as of the time of writing, not generally applicable. As a result, the application of automated scorecards is limited to automatically identified outbound deal flow and structured data. On the macro level, all micro-level SC scores are automatically aggregated with or without weightings, as is the case with manual scorecards. Automated scorecards are similarly deterministic but are more efficient than manual scorecards. In exchange for this efficiency boost, VCs need to surrender further control over both the micro and macro levels. While macro-level automation is straightforward and bears limited risk, micro-level automation is subject to two types of errors: First, I assume that with regards to the input data, incorrect information can hardly be sense-checked and thus might result in incorrect SC scores. Although algorithms are capable of identifying an increasing number of exceptions, occasionally even simple errors such as misspellings can lead them to calculate an incorrect SC score (e.g., "Stanford University" would lead to the maximum score, but "Satnford University" would lead to the minimum score). Second, there exists the possibility that the NLU models and the dictionaries used are incomplete. For example, a specific top-tier employer might not be included in a dictionary, which would result in the minimum team SC score being assigned, whereas an investment professional evaluating the situation may notice the company by chance or look it up on the internet and subsequently allocate the highest score possible. I summarize potential data and

NLU model issues as “unknown unknowns.” As it is impossible to exclude these unknown unknowns, VCs either use such models in addition to manual scorecards (i.e., create redundancy) or refrain from adopting them from the outset.

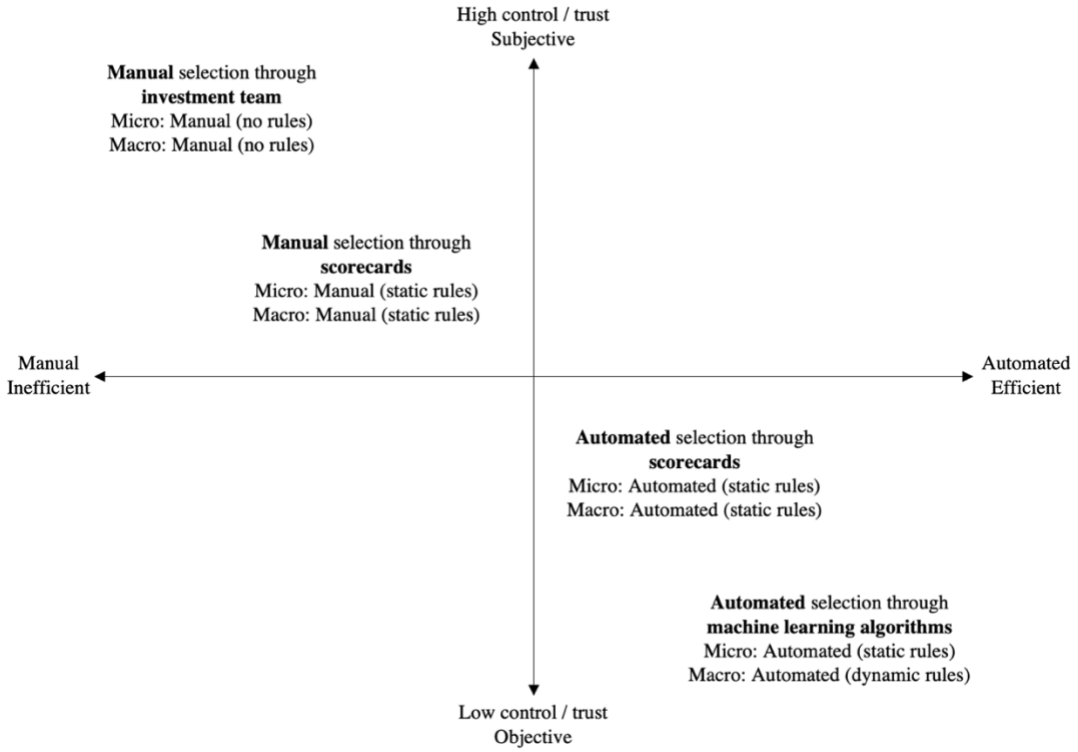
Machine learning (ML)-based selection. Both of the previously described approaches are deterministic/static across the micro and macro levels, whereas ML-based approaches are deterministic/static on the micro level but dynamic/flexible on the macro level. This means that there exists an NLU model that, similarly to automated scorecards, classifies the underlying inputs/features in a fixed structure but gradually improves the weighting of SC scores with every additional training cycle (e.g., team score becomes more important and market score becomes less important over time). Similarly to automated scorecards, the application of ML-based tools is typically limited to outbound deal flow and structured information. Due to the high degree of automation, I consider the ML-based approach the most efficient, and, in contrast to the other two approaches, there exist multiple variations of intelligent screening and success prediction tools, which have been applied in a variety of startup contexts. For example, Krishna, Agrawal, and Choudhary (2016) and Arroyo, Corea, Jimenez-Diaz, and Recio-Garcia (2019) compared the startup success prediction performance of different ML models, including random forests, decision trees and support vector machines, in a wider VC investment selection context, whereas Ghassemi, Song, and Alhanai (2020) and Catalini, Foster, and Nanda (2018) applied those models to predict startup success within a narrow entrepreneurial competition and an accelerator program, respectively. Scholars such as Hunter, Vielma, and Zaman (2016) adopted a more traditional approach by applying integer programming, a mathematical optimization technique, for startup success prediction in the context of VC portfolio construction. Although explorative studies focused on ML in VC are flourishing, the real-world application of ML-based techniques and the adoption thereof by VCs are lagging behind (Gompers et al., 2020; Schmidt, 2019). In addition to their fears concerning unknown unknowns on the micro level, VCs hesitate to surrender further control over the macro level, as tools such

as neural networks often function as black boxes, which prevents VCs from comprehending the decision-making process. The adoption of such screening tools requires VCs to fully detach and uncouple from the process.

In summary, I find that, in contrast to the automated deal *origination* or quantitative sourcing approaches, (semi)-automated *screening* approaches bear non-negligible risks, as they might mistakenly weed out successful companies (i.e., in the worst case, they may lead to a high number of FNs). As a consequence, these novel screening tools face heavy skepticism and lag behind in terms of adoption. Putting the above observations into context, I expect that the higher the degree of automation, the more efficient, scalable and objective the process, but, at the same time, the lower the investor's control, trust and adoption. I frame this phenomenon as the "*automation–control tradeoff*," which is displayed in Figure 2.3.1.

Figure 2.3.1: Automation-control trade-off

This figure displays the automation–control trade-off. The x-axis ranges from manual and inefficient approaches on the left to automated and efficient ones on the right, whereas the y-axis ranges from low control/trust and objective approaches on the bottom to high control/trust and subjective ones on the top. This figure classifies the four major groups of manual selection through investment team, manual selection through scorecards, automated selection through scorecards and automated selection through machine learning algorithms across the micro and macro levels.



I conducted 63 informal interviews with VCs between January 2018 and December 2019 to better understand practitioners' perspectives on the automation–control trade-off and ML-based screening tools. In line with my expectations, I find that VCs generally prefer to retain control over scalability, and the adoption of automated screening tools thus lags behind. Specifically, interviewees were concerned that ML-based tools would perform worse than the status quo, (i.e., deal screening via investment professionals or interns) and that unknown unknowns would remain unobserved and thus undermine their performance. Almost all interviewees were reluctant to adopt black-box tools for which they cannot comprehend the decision-making criteria. Lastly, approximately two thirds of the interviewees mentioned that they cannot allocate the resources required to implement a useful tool and plan to either rely on external providers or not use such a tool at all. Schmidt (2019) conducted 12 expert interviews focused on the adoption of AI in VC; her findings are largely in line with my own. As part of my interviews, I asked the participants about their requirements with regard to integrating white-box ML-based screening tools into their deal selection process. Ninety-five percent (or 60 out of 63 interviewees) asked for a direct comparison between such tools and investment professionals, as they were not willing to sacrifice performance compared to their status quo.

To help resolve the automation–control trade-off and to allow the VC investment process to scale, I seek to shed light on the most efficient and effective ML-based screening approaches and convince investors of the associated benefits and the limited downside risk compared to their status quo.

2.3.3 Performance benchmarking for ML-based investment screening

To credibly benchmark the performance of ML-based screening approaches against that of VC investment professionals, I start this chapter by defining the conditions and requirements for a direct comparison. Thereafter, I examine existing benchmarking approaches and select the

most suitable one for my study. Finally, I detail my setup, the company test sample and the participating group of investment professionals.

2.3.3.1 Benchmarking requirements

Benchmarking is a widely used method for comparing and evaluating the performance of different tools or processes. It is often used to compare novel approaches or tools with industry standards in a structured and replicable way (Beyer, Löwe, and Wendler, 2019). However, in contrast to most benchmarking studies, the difficulty in my setting is that I seek to compare two fundamentally different groups, namely algorithms versus humans, as opposed to algorithms versus algorithms or humans versus humans. To ensure the direct comparability of the results, I define the following four benchmarking requirement groups: (1) the same input data (i.e., the independent variables or features), (2) the same goal definition (i.e., the dependent variable or label), (3) the same boundary conditions and (4) a suitable measurement method. Overall, I need to ensure that neither party has an unfair advantage and that both parties can perform without any restrictions.

(1) Concerning the **independent variables**, I need to ensure that both parties receive the same *sufficient* input information based on which to make an *unrestricted* go/no-go screening decision. The selection of independent variables is explained in more detail below.

(2) I follow Arroyo et al. (2019) with respect to the **dependent variable** and consider a) "company closed at least one follow-on financing round," b) "trade sale" or c) "initial public offering" (IPO) as success and d) "company closed" and e) "no event" as failure. While the authors define a multi-class variable to predict the specific outcomes, I am interested in a go/no-go screening decision and hence select a binary dependent variable where 1 is attributed for either one or a combination of events a), b), and/or c) and 0 in any other case.

(3) Beyond the goal itself, it is important to consider the following **boundary conditions**:

a. First, the format of how the input information is presented to the investment professionals and the ML algorithms may vary as long as it has no direct or indirect impact on the outcome.

b. Second, I face an asymmetric cost matrix because VCs can lose their investment only once should a company go bust (i.e., a complete write-off) but can multiply their initial investment more than once should a company prove successful. This means that FNs are more costly than false positives (FPs). Most supervised learning techniques return some kind of probability $p = P(Y = 1 \mid X = x)$, which is transformed into binary predictions using – in the standard case of symmetric cost learning – $Y = 1$ if $p >= 0.5$, meaning that if $Y = 1$ is more likely than $Y = 0$, then predict 1. The issue with cost-sensitive learning is that the threshold of $p = 0.5$ is no longer optimal. Instead, if FNs are more costly than FPs, I need to lower the probability threshold so that I already predict $Y = 1$ when $P(Y = 1 \mid X = x) = 0.25$, for example, meaning that due to the high cost of classifying a successful company as a failure, I already predict 1 with less confidence. At the same time, the probability threshold directly determines how many selected items or predictions of the relevant class (= 1 / S) the model returns. Both metrics are inseparably connected (i.e., the lower the probability threshold p, the higher the number of predicted items S and vice versa). For example, in the extreme case of $p = 0$, the model would select all items of the relevant group and thus reduce the number of FNs. Although this would be an optimal outcome in terms of FNs, it would be useless in my setting, as the screening algorithm would qualify all companies for further evaluation. While, in

theory, I could either define *p* and have *p* determine *S* or define *S* and have *S* determine *p*, the reality that there is a limited period of time available for further deal evaluation and thus a limited number of selected opportunities allows only the latter. Most VCs can estimate approximately how many hours within a specific period of time they can allocate to evaluating new investment opportunities in more detail (e.g., 10 hours per week). Assuming an on-average constant evaluation time per company, the number of companies that can be further assessed within a given period of time will be constant as well. Hence, I require a fixed number of selected items *S* as a result of the screening process, which then determines the probability threshold *p*, and seek to reduce the number of FNs as much as possible.

c. Third, I must ensure reliability and validity by maximizing the sample size of participants.

(4) Most importantly, the outcomes of both parties need to be directly comparable. In line with the extensive ML and decision-making literature, I rely on a confusion matrix and its underlying metrics for a detailed performance measurement (Arroyo et al., 2019; Gastaud, Carniel, and Dalle, 2019; Ghassemi et al., 2020; Sokolova, Japkowicz, and Szpakowicz, 2006).

a. Although accuracy (AC)[23] is helpful to determine the overall prediction performance, reducing the number of FNs and thus increasing sensitivity or recall (RE)[24] at a fixed number of positive predictions *S* is of utmost importance in my setting. I agree with Arroyo et al. (2019)'s statement that *'Recall is not a critical measure. VCs do not need to find all the*

[23] Accuracy (AC) is calculated by dividing the sum of true positives (TPs) and true negatives (TNs) by the total number of predictions.
[24] Recall (RE) indicates "how many relevant items are selected" and is calculated by dividing the number of TPs by the sum of TPs and FNs.

"interesting" companies available in the world (or in the database) since they cannot invest or even consider investing in all of them' for the final investment decision stage but strongly disagree regarding the general screening stage. Given a fixed number of selected companies *S*, I assume that maximizing RE and consequently reducing the risk that a deselected company might prove successful is the most important dimension to consider.

b. The resolution of AC and RE depend on the respective denominator (i.e., for AC, it depends on the total number of predictions, whereas, for RE, it depends on the number of actually positive outcomes).[25] Consequently, I must increase the company sample size/number of predictions and the number of successful companies to increase the degree of detail in my performance comparison.

2.3.3.2 Benchmarking approaches

There exist a range of well-established benchmarking approaches for comparing humans with algorithms (Esparza, Scherer, Brechmann, and Schwenker, 2012; Jiang, Ye, Chang, Ellis, and Loui, 2011; Stallkamp, Schlipsing, Salmen, and Igel, 2012), but few have been applied in startup selection contexts. The performance evaluation for the ML algorithms is straightforward, as I feed the models with structured input data and summarize their predictions in a confusion matrix. In contrast, the determination of investment professionals' screening performance is more complex and can be approached via different methods.

In a field study, for example, I might collect real-world screening outcomes (i.e., go/no-go decisions after initial screening) for specific companies as documented in VCs' notes or

[25] For example, if a company sample includes 50 companies with 10 successful and 40 unsuccessful ones, the AC resolution will equal 2% per increment (1/50), whereas the RE resolution will equal 10% per increment (1/10). To properly benchmark the RE performance between VCs and the ML algorithm in this example, either party needs to be at least 10% better than the other, as, otherwise, there would be no visible difference.

customer relationship management systems. Collecting field data would likely lead to a large and diversified sample, as the related costs of participation for a venture capitalist would be minimized (assuming that they are willing to share such information). However, all collected screening decisions would have been made based on different input data, with different goals in mind and under varying circumstances. Such inconsistency would clash with almost all of my benchmarking requirements and would thus be unsuitable.

Alternatively, I might conduct a controlled experiment in which I ensure that all VCs have the same input information, goal definition and circumstances. For example, Ghassemi et al. (2020) compared startup competition results with ML algorithm performance by providing the same input data to the competition's judges and the algorithms, and ensuring that all participants were aligned in terms of goals and circumstances. While such offline experiments would fulfill benchmarking requirements (1), (2) and (4), the required physical presence and fixed time allocation would increase the barriers to participation for VCs, which would in turn drastically decrease the participant sample size and potentially the diversity in terms of participants' geographic locations, which would violate my third boundary condition (3c). Online experiments, and more specifically online surveys, mitigate this limitation, as they allow VCs to individually select the most suitable time to participate and eliminate the need to be physically present. Thus, the use of an online survey would ensure that, regardless of geographical and time constraints, every VC could theoretically participate, which would in turn increase the participant sample size and geographic diversification. Given that an online survey best suits my benchmarking requirements, I select this approach for my study.

2.3.3.3 Survey design

Prior studies on survey response rates suggest that response rate mainly depends on the topic and length of a survey. I assume that the topic of my study directly interests the respondents and thus has high salience, which leaves the length of the survey as the remaining factor. Although some studies have shown that no relationship exists between survey length

and response rate (Bruvold and Comer, 1988; W. S. Mason, Dressel, and Bain, 1961), the majority of scholars have found a negative correlation between both factors (Cook, Heath, and Thompson, 2000; Edwards et al., 2002; Heberlein and Baumgartner, 1978; Singer, 1978; Walston, Lissitz, and Rudner, 2006; Yammarino, Skinner, and Childers, 1991). Consequently, I seek to reduce the survey duration to increase the participant sample size and thus ensure (3c). This approach in turn forces us to reduce either the number of companies included (which might clash with (4b)) or the number of SC included (which might clash with (1)).

To resolve this three-dimensional trade-off, I define two of the three parameters to obtain the third. I assume that if I include enough actually successful companies in my sample to achieve a sufficient resolution in terms of AC and RE (as required by (4b)) and if I ensure that neither the VC nor the ML algorithm sacrifice performance due to omitted SC information (as required by (1)), maximizing the completion rate/participant sample size (as required by (3c)) are of primary importance. Hence, I followed the dual process below to identify the required screening time per company based on the optimal number of SC to be included before selecting the ideal survey duration and dividing it by the required screening time per company to obtain the number of companies to be included.

1) *Ensure that most important VC SC are included (described in more detail below)*

- Extract ranking from most to least important SC for VCs from the literature.
- Run pre-tests with VCs to identify the best balance between the number of SC included and the screening time required by gradually including SC on random companies and measuring the required screening time as well as the perceived confidence of being able to make a go/no-go screening decision; based on the results, create a list of SC to be included in the test dataset (*VC SC list*).
- Ask VCs after they have successfully completed the online survey whether they would have required additional information on either one of the companies to make an

informed go/no-go screening decision; exclude all VCs who required further information.

2) *Ensure that most important ML SC are included (described in more detail in Section 2.3.4)*

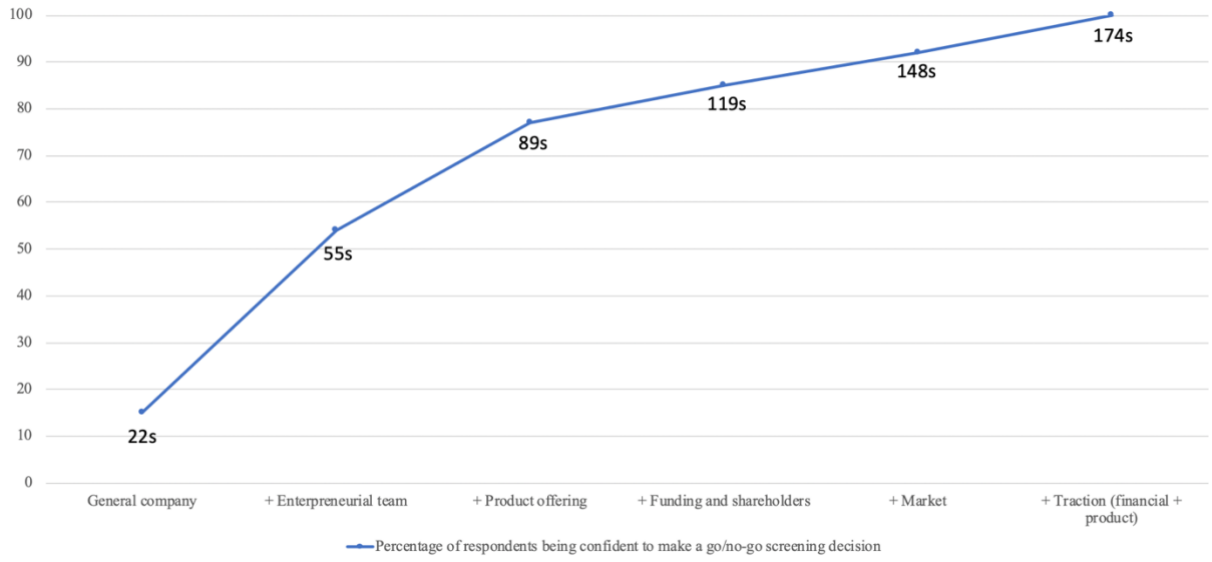
- Train and select the best-performing ML algorithm based on all available SC (*full SC list*) and rank SC by importance based on their respective weights.
- If not already included, add the most important ML SC from the *full SC list* to the *VC SC list* as long as the impact of doing so on required screening time is limited; this results in the *test SC list.*
- Train the selected ML algorithm based on the final *test SC list* and compare the performance thereof to that of the algorithm based on the *full SC list* to quantify a potential performance drop due to omitted SC; this potential drop needs to be considered when ultimately interpreting the benchmarking results.

I started with the first branch of the dual process and ranked all SC groups by their respective importance for the overall screening decision based on my previous literature research. Next, I incrementally added SC information to a company one-pager, beginning with the most important. As the "general company" SC group contains data on headquarters, industry and a short description of each business, it is required to be examined with respect to the VC's hard SC and thus needed to be inevitably included. In line with the results of my literature review, Gompers et al. (2020) found that team, product and fit with a venture capitalist's investment focus are considered the most important SC, whereas valuation, market and financial metrics are less relevant. I followed this order when adding the SC to my one-pager. I then asked 13 random European VCs to analyze a different random company with an increasing number of SC included per participant and to decide whether they would be able to make a go/no-go screening decision. For example, VC#1 screens company 1 with one SC, VC#1 screens company 1 with two SC, VC#1 screens company 1 with three SC, etc., whereas VC#2 screens company 2 with one SC, VC#2 screens company 2 with two SC, etc. Lastly, I recorded the time

required for screening the companies. The purpose of this pre-test was to ensure that I identify the optimal balance between SC included and the screening time required per company to confidently make a go/no-go decision, independent of the underlying company or the evaluating venture capitalist. Figure 2.3.2 exhibits the SC included on the x-axis: From left to right, I added one SC per increment to the sum of the ones on the left, and the average percentage of participants who were confident in making a go/no-go screening decision per number of SC criteria based on the provided information. In addition, the figure depicts the average screening time required for the sum of all SC included up until the respective stage in seconds (*s*).

Figure 2.3.2: Percentage of VCs confident to make decision based on provided SC

The figure displays the percentage of VCs who are confident in making a go/no-go screening decision based on the provided SC. The x-axis shows the included SC. From left to right, I add additional SC to the ones on the left. The y-axis shows the average percentage of VCs who were confident in making a go/no-go screening decision based on the provided SC. In addition, the figure shows the average time required to read through the provided SC in seconds (s).



In line with the extensive literature on the importance of entrepreneurial teams, I find a confidence increase from 15.4% (two out of 13) based on the general company information alone to 53.8% (seven out of 13) when team information was added to the one-pagers. Similarly, I observe two step increases from 53.8% to 69.2% (nine out of 13) when further adding product

information and from 69.2% to 84.6% (11 out of 13) when further adding funding and shareholder information. Market- and traction-related information each account for 7.7%, which supports my assumption that they are the least relevant SC. Taken together, the four SC general company information, entrepreneurial team, product offering and funding and shareholder information lead to an average go/no-go confidence level of approximately 85% at an average screening time per company of 119 seconds or approximately two minutes. Adding additional SC such as market or traction only slightly increases the perceived confidence level while significantly increasing screening time in relative terms. Zacharakis and Meyer (2000) support reducing the basis of information on which VCs make decisions to improve selection quality, as they find that *"With more information, the accuracy of [a venture capitalist's] decision remains unchanged and may even decrease. Part of the reason that VCs' accuracy is not greatly improved with more information is that experts, despite their beliefs, use relatively few available cues. Therefore, VCs tend to mistakenly believe that they are making a more informed decision with a greater amount of information even though they are likely ignoring the additional information or using it inappropriately."* Åstebro and Elhedhli (2006) provide additional evidence that more information leads to worse decisions of VCs. Therefore, the choice to include the four above-mentioned SC groups on my *VC SC list* is intended to not only facilitate making accurate predictions and optimize the screening time to approximately two minutes per company but also to increase the efficiency of the decision-making process.

As mentioned previously, I included the question of whether the respondent would have required additional information to make a go/no-go screening decision at the end of the online survey to ensure that the omission of further SC has no perceived negative impact on the participant's screening performance. To also consider the ML perspective, I added individual SC from the *full SC list* that are important for the ML algorithm but were not yet included in the *VC SC list*. Due to the underlying complexity of the ML algorithm training, I dedicate Chapter 4 to this topic and only consider the resulting SC to be added at this point, namely

founder_age. As a result, I obtain the final *test SC list*, which fully satisfies benchmarking requirement (1). The expected screening duration per company remains unchanged at approximately two minutes.

Next, I sought to identify the ideal online survey duration for VCs to maximize the completion rate and the participants sample as required by (3c). Revilla and Ochoa (2017) found that the optimal survey length is a median of 10 minutes and that the maximum survey length is 20 minutes. Longer surveys significantly reduce the completion rate. Given that it takes approximately two minutes to screen one company based on my *test SC list*, a 10-minute survey would result in a company sample size of five and thus a maximum AC resolution of 20% per increment, whereas a 20-minute survey would result in a company sample size of 10 and thus a maximum AC resolution of 10% per increment (as described in (4b)). Although it would be desirable to obtain 5% or even less per increment, doing so would require us to at least double the survey duration to 40 minutes or more. As such a duration would significantly reduce the completion rate, I selected 20 minutes and 10 companies to be included as a compromise. Knowing that the success/failure ratio for early-stage ventures is approximately 1:5 (see Section 2.3.3.4 below), my rather small sample size of 10 companies, together with benchmarking requirement (4b), prevented me from transferring such imbalanced ratios to my dataset. For example, a sample of 10 companies with two successful and eight unsuccessful ones would result in a RE resolution of 50% per increment (1/2), which is clearly not sufficient. To optimize the resolution and ensure a proper performance comparison, I selected a balanced dataset consisting of five successful and five unsuccessful companies, resulting in a resolution of 20% per increment for both AC and RE (1/5). As (3b) requires me to explicitly communicate the number of successful companies $S = 5$ to be selected, I managed expectations accordingly and hence had no issue amending the underlying success/failure ratio.

Finally, I designed my online survey in Qualtrics based on a 20-minute target duration and included 10 companies (five successful and five unsuccessful), with both general company

information as well as team, product, funding and shareholder information being provided for each. I conducted the survey throughout January and February 2020.

2.3.3.4 Company sample

Similar to VCs' approach of narrowing their investment strategies across the three dimensions of geography, industry and stage (Retterath and Kavadias, 2020), I defined a specific focus for my dataset to accurately target the venture capitalist participant group. By doing so, I sought to increase the completion rate and validity of my online survey. I focused my company test sample on European software companies founded after January 1st, 2010 (founding cutoff t_f) that had successfully raised a seed-financing round at a flexible input point in time t_i between January 1st, 2015 and December 31st, 2016. The flexible input point allowed me to ensure that all company SC are represented at the same development status (i.e., directly after their seed-financing rounds). I collected the independent variables as of t_i and the dependent variable as of December 31st, 2019 (output point in time t_o), between three to five years after the seed-financing round. In line with Arroyo et al. (2019), I define the period between t_f and t_i as the warm-up window and the period between t_i and t_o as the simulation window.

To ensure reliable and valid performance of the participating investment professionals and a potential real-world application of the ML screening algorithms, I created my company sample based on actual company information provided by a variety of VC data providers. As there exists no comprehensive and perfect quality startup database (Kaplan and Lerner, 2016; Kaplan et al., 2002; Lerner, 1995; Maats et al., 2011; Retterath and Braun, 2020), I decided to assemble my dataset based on data obtained from various providers. In line with Retterath and Braun (2020), I would have theoretically collected all general company and product information from VentureSource, founder- and team-related information from Pitchbook and/or Crunchbase and funding-/shareholder-related information from a combination of VentureSource, Crunchbase and Pitchbook. However, Dow Jones, the operator of VentureSource, discontinued

the service as of March 31, 2020. Moreover, Pitchbook does not offer a bulk export function and limits the maximum number of downloadable companies per month to a thousand. Clearly, this number would not have been sufficient in terms of sample size. As a consequence, I followed Arroyo et al. (2019) and used Crunchbase data as my foundation but complemented specific SC and verified these SC via automated searches and fuzzy matching in Pitchbook and LinkedIn.

The resulting *full dataset* of European software companies that were founded after January 1st, 2010 includes 77,279 organizations with 118,231 verified founders. Based on the above-mentioned success definition, 14,142 companies or 18.3% of the dataset are considered successes (dependent variable = 1), whereas 62,137 companies or 81.7% of the dataset are considered failures (dependent variable = 0). These imbalanced figures are approximately in line with Arroyo et al. (2019). While working with actual company data ensures real-world application and the external validity of my performance benchmarking, it poses the risk that the surveyed VCs may recognize the underlying companies based on the SC information as of t_i and already know the actual success/failure outcomes as of t_o. Although I cannot fully eliminate the risk of this potentially unfair advantage, I mitigate it by only including less prominent companies and anonymizing obvious information such as company, founder or product names. Again, I asked the 13 VCs from my pre-test whether they would be able to identify specific companies from my *full dataset*. Subsequently, I randomly selected 10 companies, five of which are considered a success and five of which are considered a failure, that were not identified by the VCs and defined them as my *test dataset*. I removed the 10 companies that were included in my *test dataset* from the *full dataset* and obtained the *training dataset*. In addition to only including less prominent companies and anonymizing the data, I ensured that the surveyed VCs have not identified any of the companies in my test dataset by explicitly asking them at the end of the online survey whether they had identified one or more of the

companies. Should the answer be "yes," I excluded the VC in question from my analysis due to an unfair advantage over the ML algorithm.

With regards to the dataset format, I fed the ML model with a structured table where each line represents a unique company and each column represents information on a specific SC. For the investment professionals, I created a one-pager per company representing all of the SC in a structured form. I assume that the different formats do not provide either party with an unfair advantage.

2.3.3.5 Participants sample

Based on the geographic, industrial and stage focus of my dataset, I reached out to 500 VCs who claim on their respective websites that they have a track record of investing in European early-stage software companies. I contacted 72 VCs from my personal network, 34 via warm introductions by mutual contacts, 168 via e-mail addresses identified with Hunter[26] and 226 via LinkedIn. One hundred and forty-eight VCs (29.6%) completed the online survey, but a total of 37 respondents had to be excluded, which resulted in a viable respondent sample of 111 VCs. Thirty-three respondents or 22% were excluded, as they requested additional SC information to make an informed go/no-go decision. Although higher than my expectation of approximately 15% based on the pre-test, the result is in the same ballpark and leaves a sufficiently large participant sample size for my analysis. The excluded VCs asked for additional information on traction (28 out of 33), market size and growth (13 out of 33), business model (seven out of 33), product differentiation and intellectual property (six out of 33) and additional financial information (three out of 33); multiple requests were allowed. Moreover, two respondents needed to be excluded, as they mentioned that they have identified one or more of the companies in my test sample, and another two needed to be excluded because

[26] "*Hunter lets you find email addresses in seconds and connect with the people that matter for your business."* Description taken from the company's website as of April 24, 2020.

they did not identify software as their industrial focus area. Table 2.3.3 characterizes the respondent sample.

Table 2.3.3: Characteristics of online survey respondents

Overview of survey respondents distinguished by headquarter country, investor type, assets under management (AUM), industry focus, years of VC experience, highest academic degree and area of highest academic degree. A total of 148 VCs completed the survey, but 37 respondents need to be excluded due to several reasons resulting in a final respondent sample 111 VCs.

Characteristic	N	% total	Characteristic	N	% total
Headquarter country	111	100	Industry focus (multiple selection)	250	100
Austria	4	4	Software/IT	111	44
France	21	19	Biotech/Healthcare	51	20
Germany	36	32	Hardware	45	18
Luxemburg	3	3	Others	43	17
Netherlands	4	4	Years of VC experience	111	100
Sweden	3	3	x < 5	57	51
Switzerland	5	5	5=< x < 10	30	27
United Kingdom	25	23	10=< x < 15	12	11
Others (<2 respondents)	10	9	15=< x	12	11
Investor type	111	100	Highest academic degree	111	100
Institutional VC	92	83	Ph.D.	7	6
Corporate VC	4	4	MBA	17	15
Family Office	5	5	Master	59	53
Other	10	9	Bachelor	26	23
Assets under management (in € millions)	111	100	None	2	2
<25	18	16	Area of highest degree	111	100
25=< x < 100	28	25	Science, Technology, Engineering, Math (STEM)	25	23
100=< x < 500	41	37	Business	72	65
500=< x	24	22	Others	14	13

Most respondents are headquartered in Germany (32%), the United Kingdom (23%) and France (19%). The rest are almost equally distributed across Europe. Eighty-three percent of the respondents consider themselves as institutional VCs with dedicated funds, whereas 4% are CVCs and 5% are family offices. As I reached out to early-stage VCs, it is not surprising that 41% of the funds have less than €100 million in assets under management (AUM) and that only 22% have more than €500 million in AUM. While 100% of the respondents focus on software/information technology, some also invest in biotech/healthcare (46%), hardware (41%) or other areas (39%). Approximately half of the respondents have less than five years' VC experience, whereas 27% have between 5 to 10 years, and another 22% have more than 10 years. I find that 75% have at least one higher degree (Ph.D., MBA or master's) and that only 2% of respondents have no degree. Unsurprisingly, 65% of respondents have business

backgrounds, and only 23% have their highest degree in a STEM field. Some of the participants who agreed to participate and to be disclosed work for firms such as 24Haymarket, Acton Capital, Amadeus Capital, Axon Partners, Bayern Kapital, b-to-v Partners, Creandum, Earlybird Venture Capital, Forward Partners, Frog Capital, Global Founders Capital, Heartcore Capital, Holtzbrinck Ventures, Idinvest Partners, Investiere, IQ Capital, Lifeline Ventures, Partech, Peak Capital, Prime Ventures, Project-A, Senovo, Speedinvest and UnternehmerTUM Venture Capital. Next, I elaborate on the training of my ML algorithm.

2.3.4 Training an ML algorithm for VC investment screening

I explain in detail below how I trained a variety of supervised classification models and selected the most suitable one for my screening performance benchmarking with investment professionals. To ensure that the most important SC groups required by the final ML algorithm are included in the test dataset, I ranked the individual SC based on the trained weights and complemented the *VC SC* list accordingly. Moreover, I trained the model based on all available SC (*full SC* list) and compared its performance with that of the model based on the *test SC* list. Hereby, I was able to quantify a potential performance drop due to omitted variables which would need to be considered when interpreting the benchmarking results.

2.3.4.1 Data preparation

I started the data preparation process by codifying my sole dependent variable. I assigned 1 to the label of my dataset should a company have been acquired, gone public or raised at least one subsequent financing round between t_i and t_o, and 0 in any other case. The process was more complex for independent variables, as their data type ranges from numerical variables such as "*total amount of capital raised*" through categorical variables such as "*industry*" (software, healthcare, financial, industrial, etc.) to binary variables such as "*social media profile available*" (yes/no). Overall, there were 163 features available in my *full dataset*. After removing unnecessary variables and cleaning missing values, I obtained a set of 45

features. To avoid assuming the existence of meaningful relationships between categorical variables, I applied one-hot encoding and transferred the categorical variables into binary variables. For example, one-hot encoding the categorical variable industry, which consists of 15 different sectors to choose from, would result in 15 individual binary variables. Each variable represents one sector and can be either 0 or 1. One-hot encoding 13 categorical features from the overall set of 45 features resulted in a total of 234 features (*full SC list*).

2.3.4.2 Model building and selection

As described in the discussion of the boundary conditions in Section 2.3.3, I face an asymmetric cost matrix where FNs are more costly than FPs. Due to the limited human resources available for analyzing selected items in more detail, the ML algorithm needs to predict a fixed number of items *S* with a flexible probability threshold *p*. Moreover, my setting requires a white-box technique that allows us to determine the relative importance of the different SC to ensure that I have included the most important ones in my test set. Following Arroyo et al. (2019) and Ghassemi et al. (2020), I considered supervised classification techniques, including decision trees (DT), random forests (RFs), gradient boosted trees (also known as "XGBoost" (XG)) and naive Bayes (NB), and added deep learning (DL) models, generalized linear (GL) models and logistic regressions (LRs) to the selection. As a result, I covered the most relevant ML classification models.

I initially trained all seven models based on the same training dataset, which includes the *full SC list*, and compared their screening performance. As described above, I selected the best-performing algorithm based on AC and RE. Table 2.3.4 shows the performance of all seven models. I selected the XG algorithm, as it clearly outperforms all other models based on both metrics. To allow the reader to understand the XG model in more detail, Figure 2.3.4 presents the relative importance of the codified SC for the overall go/no-go prediction (i.e., the *full SC list*).

Table 2.3.4: Screening performance comparison of ML algorithms

Performance overview of Decision Trees (DT), Deep Learning Models (DL), Generalized Linear Models (GL), Gradient Boosted Trees (also known as "XGBoost models", (XG)), Logistic Regressions (LR), Naive Bayes Models (NB) and Random Forests (RF). All models are trained based on our training dataset and all available selection criteria (SC) information, i.e. the full SC list. Performance is compared based on global model accuracy (AC) and recall (RE).

Machine learning technique	AC	RE
Decision Trees (DT)	70.4	82.5
Deep Learning (DL)	80.3	82.8
Generalized Linear Models (GL)	81.7	80.5
Gradient Boosted Trees (XG)	82.6	83.5
Logistic Regressions (LR)	68.6	73.6
Naive Bayes (NB)	76.4	78.2
Random Forests (RF)	63.8	76.4

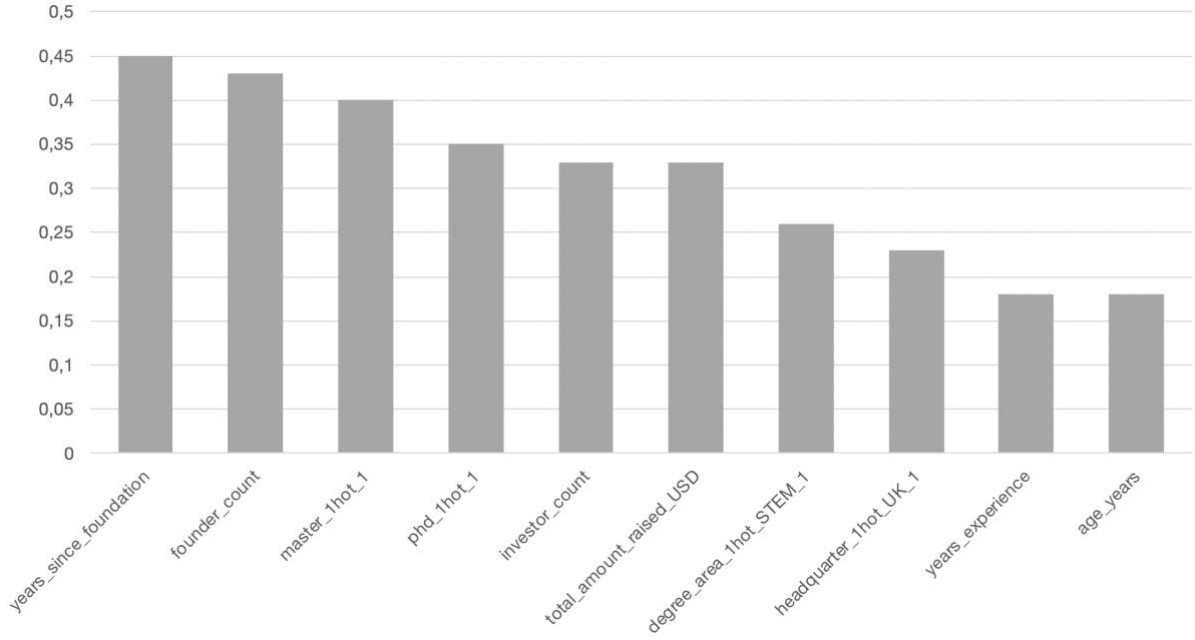
Table 2.3.5: Screening performance comparison of XGBoost algorithm based on different selection criteria

Performance overview of Gradient Boosted Trees (also known as "XGBoost models", (XG)) trained based on our training dataset and all available selection criteria (SC) information, i.e. the full SC list, and based on our training dataset and the reduced SC list, i.e. the test SC list. Performance is compared based on global model accuracy (AC) and recall (RE).

Selection criteria (SC) list	AC	RE
full SC list	82.6	83.5
test SC list	80.3	81.4

Figure 2.3.3: Feature importance for XGBoost algorithm

The figure exhibits importance of each feature for the XGBoost algorithm based on the training dataset and the full SC list. The x-axis shows the 10 most relevant features, with the most important one on the left and the least important one on the right, whereas the y-axis displays the relative weights of each feature.



I find that *years_since_foundation* has the strongest predictive power (or macro-SC importance), closely followed by *founder_count* and *master_1hot_1*, the latter of which indicates whether the CEO of the company has a master's degree. Matching the top 10 items of the *full SC list* with the *VC SC list* shows that all metrics (with the exception of *age_years*, which represents the age of the founding CEO) were already explicitly or implicitly (e.g., *years_since_foundation* is implicitly included as founding year) included in the *VC SC list*. Although this was not the case for several SC in the long tail of the *full SC list*, their limited predictive power prevented me from including them in the *test SC list*. Therefore, I only added the age of the CEO to the *VC SC list*, resulting in the *test SC list* as described in the discussion of the company sample in Section 2.3.3. Next, I retrained the XG algorithm based on the training dataset and the reduced *test SC list*. While the *full SC list* comprises 234 features, the *test SC list* contains 135 features. Table 2.3.5 shows the performance comparison of the XG algorithm based on the *full SC list* and on the *test SC list*. While the algorithm based on the full

SC list achieves 82.6% in AC and 83.5% in RE, the algorithm trained with the test SC list performs only slightly worse, achieving 80.3% in AC and 81.4% in RE. I selected the XG algorithm based on the training dataset with the *test SC list* as my final benchmarking algorithm.

Table 2.3.6: Screening performance benchmarking of VC investment professionals and XGBoost algorithm

Performance overview of VC investment professionals online survey results and Gradient Boosted Trees (also known as "XGBoost models", (XG)) trained based on our training dataset and the test SC list. We compare the screening performance based on global model accuracy (AC) and recall (RE). Note that due to the balanced nature of the test dataset (5 successful and 5 unsuccessful companies) and the fixed number of selected items S (both parties were required to select exactly 5 companies for further evaluation), AC and RE are equal and thus we do not distinguish in this table. The table shows the number of predictions (N), the percentage of the total within a specific group (% total), the minimum percentage value for AC and RE (Min %), the maximum percentage value for AC and RE (Max %), the median percentage value for AC and RE (Median %) and the average percentage value for AC and RE (Avg. %).

Benchmarking group	N	% total	Min %	Max %	Median %	Avg. %
VC investment professionals	1110	100	20	80	60	57
Investor type	1100	100	20	80	60	57
Institutional VC	920	83	20	80	60	58
Corporate VC	40	4	40	60	40	45
Family office	50	5	40	60	60	56
Others	100	9	20	80	60	56
Asstes under management (AUM)	1110	100	20	80	60	57
<25	180	16	20	60	40	51
25=< x < 100	280	25	40	80	60	60
100=< x < 500	410	37	20	80	60	54
500=< x	240	22	20	80	60	62
Years of VC experience	1110	100	20	80	60	57
x < 5	570	51	20	80	60	58
5=< x < 10	300	27	40	80	60	58
10=< x < 15	120	11	20	60	60	54
15=< x	120	11	20	60	60	52
Highest academic degree	1110	100	20	80	60	57
Ph.D.	70	6	40	80	60	61
MBA	170	15	20	80	60	57
Master	590	53	20	80	60	57
Bachelor	260	23	40	80	60	56
None	20	2	40	80	40	48
Area of highest academic degree	1110	100	20	80	60	57
Science, Technology, Engineering, Math (STEM)	250	23	40	80	60	61
Business	720	65	20	80	60	57
Others	140	13	40	80	40	50
Gradient Boosted Trees (XGBoost, "XG")	10	100	80	80	80	80

2.3.5 Benchmarking results: ML algorithm versus VC investment professionals

I conducted the online survey and ran the trained XG algorithm based on the test dataset and the *test SC list*. Neither party had seen or analyzed the information contained in the test dataset previously. Table 2.3.6 summarizes the results. The table shows the number of predictions (*N*), the percentage of the total within a specific group (*% total*), the minimum percentage value for AC and RE (*Min %*), the maximum percentage value for AC and RE (*Max*

%), the median percentage value for AC and RE (*Median %*) and the average percentage value for AC and RE (*Avg. %*). Due to the balanced nature of my test dataset (which consists of five successful and five unsuccessful companies) and the requirement that exactly five companies be selected for further evaluation ($S = 5$), AC and RE are always equal, and the displayed percentages thus represent both metrics. As described in the "Survey design" part of Section 2.3.3, the resolution for all *Min*, *Max* and *Median* metrics is 20% per increment, whereas it is 1/555 for the *Avg.*[27] values of all VC investment professional groups. In contrast, the resolution per increment for all AC and RE metrics of the XG algorithm is 20%. Each 20% increment in AC and RE represents one TP and one TN. For example, 40% in AC and RE means that two actually successful companies were predicted to be successes (TPs) and two actually unsuccessful companies were predicted to be failures (TNs), whereas three actually successful companies were predicted to be failures (FNs) and three actually unsuccessful companies were predicted to be successes (FPs).

I divide the VC investment professional group into subgroups based on the same characteristics as depicted in Table 2.3.3 to identify characteristic-specific performance dependencies. My results show that VC investment professionals perform between 20–80% and that the median of 60% is slightly greater than the average of 57%, indicating a left-skewed distribution and some outliers on the lower end. In terms of investor type, institutional VCs perform best with a median of 60% and an average of 58%, whereas CVCs perform worst with a median of 40% and an average of 45%, making for relative differences of 33% and 22% between the subgroups. It is important to note that institutional VCs account for 83% of the respondents, while corporate VCs only account for 4%. Although the results show similar dependencies for different AUM groups, I cannot find a consistent correlation between the amount of AUM and the average performance. Investment professionals from firms with more than €500 million in AUM achieve a median of 60% and an average of 62%, whereas

[27] Note that of 1,110 predictions, 555 companies are actually successful due to the balanced nature of our test dataset.

investment professionals from firms with less than €25 million in AUM achieve a median of 40% and an average of 51%, making for relative differences of 33% and 18% between the subgroups on both ends of the spectrum. In between, however, the performance increases from the first to the second AUM group, decreases from the second to the third AUM group and then increases again. This is not the case for VC experience and level of education, however. While years of VC experience do not seem to impact the median of 60%, the average is 58% for VCs with less than 10 years of experience and decreases to 54% by year 15 and even to 52% thereafter. This represents a relative decrease in average performance of 12%, meaning that after a period of approximately 10 years in VC, the more years of VC experience an individual has, the worse the screening performance becomes. In terms of VC education, I find a positive correlation between the level of the highest academic degree attained and the average performance. While VCs without a degree achieve a median of 40% and an average of 48%, the median jumps to 60% and the average to 56% for VCs with a bachelor's degree. The median stays constant at 60% for VCs with master's, MBA and Ph.D. degrees, but the averages further increase to 57%, 57% and 61%, respectively. These figures represent a relative increase from the lowest to the highest value of 33% for the median and 21% for the average, meaning that the higher the academic degree, the better a venture capitalist's screening performance. Concerning the field in which the highest degree was earned, I find that a STEM background leads to a median of 60% and an average of 61%, whereas, for business backgrounds, the median stays the same but the average drops to 57%. For all other backgrounds, the median decreases to 40% and the average drops to 50%, representing a relative difference of 33% for the median and 18% for the average between "others" and STEM backgrounds.

The performance of the XG algorithm based on the test dataset and the *test SC list* as presented in Table 2.3.6 is approximately in line with the performance of the training dataset and the *test SC list* presented in Table 2.3.5. AC and RE are 80% meaning that only one in five actually successful companies was predicted to be a failure. Comparing the performance of the

ML algorithm with that of the VC investment professionals, I find that the XG algorithm outperforms all investment professionals by 20% (ML = 80% vs. VC = 60%) in terms of median AC and RE values and by 23% (ML = 80% vs VC = 57%) in terms of average AC and RE values. In relative terms, the XG algorithm performs 25% better with respect to median values and 29% better with regard to average values. As the resolution for the maximum AC and RE values is set at 20% per increment, it is difficult to draw detailed conclusions on whether the XG algorithm performs slightly worse than, exactly equal to or better than the best-performing VC investment professionals. Based on the results and resolution of my comparison, I can confidently state that the XG algorithm performs approximately as well as the best-performing VC investment professionals in my study.

2.3.6 Discussion and implications

In this paper, I conducted an investment screening performance benchmarking between 111 VC investment professionals and a supervised XG classification algorithm to create trust in ML-based screening approaches, accelerate their adoption and ultimately enable the traditional VC model to scale. I provided anonymized company information on 10 European early-stage software startups, including data on the entrepreneurial teams, products, funding situations and shareholder structures, via an online survey to 111 VCs and asked them to predict the success versus failure outcome of the included companies. As the input information had been collected throughout 2015 and 2016 immediately following the completion of each startup's seed-financing round, I know that five companies in my sample had proven successful and five had proven unsuccessful as of January 2020, the point in time at which I conducted my survey. Thus, I was able to compute a confusion matrix and analyze the prediction performance. In parallel, I trained a variety of ML algorithms based on a wider dataset, selected the XG classifier as the best-performing model, provided it with the same company information as the investment professionals and collected the predictions from both the professionals and

the algorithm concerning the success or failure of each startup. Finally, I compared the screening performance based on AC and RE across the different subgroups of investment professionals; in addition, and more importantly for my study, I compared the performance of the investment professionals and the XG algorithm.

2.3.6.1 Implications

My study makes contributions to two broad areas: First, my results contribute to the growing literature on the use of AI/ML in VC (Arroyo et al., 2019; Catalini et al., 2018; Ghassemi et al., 2020; Schmidt, 2019) and help academics and VC practitioners to better understand the performance of ML-based screening tools compared to that of the status quo. Although I find that my XG classification algorithm performs relatively 25% better than the *median* venture capitalist and 29% better than the *average* venture capitalist in screening and selecting European early-stage software companies, by no means do I suggest replacing humans in the screening stage. Instead, I recommend an augmented approach where ML-based screening tools narrow the upper – steadily growing – part of the deal funnel to a constant number of investment opportunities, which can then be double-checked and further evaluated by investment professionals. Using this approach, new investment opportunities can be selected in an objective and highly efficient way, and, as a result, investment professionals could save substantial time that could then be focused on properly evaluating a selection of high-potential deals. Moreover, they can use the freed-up resources to build stronger relationships with the selected entrepreneurial teams and to put themselves into a better position to secure the most competitive deals. Instead of going broad and shallow by allocating limited resources to an ever-growing number of opportunities, the use of ML-based screening tools frees up time and allows a venture capitalist to go narrow and deep on a selected number of opportunities while still ensuring that promising deals are not overlooked. Ultimately, this data-driven approach helps VCs to scale their traditional operations and remove potential biases in the screening

process (Franke et al., 2006; Paul A Gompers et al., 2020; Zacharakis and Meyer, 1998; Zacharakis and Shepherd, 2001).

Second, my results contribute to the rich literature concerning the VC investment process and specifically that on the generic screening stage (Fried and Hisrich, 1994) by presenting a comprehensive characteristic-specific performance analysis for VC investment professionals. My findings indicate that institutional VCs perform better in the generic screen stage than CVCs. It may be the case that CVCs typically evaluate opportunities with a different – potentially more strategic – lens than institutional VCs, which purely focus on outsized returns. As a consequence, CVCs might lack the skills required to screen deals based purely on an exit-focused perspective. Additionally, I find that after about a decade of VC experience, there exists a negative correlation between experience and screening performance. Although this finding might initially seem counterintuitive, it can be explained with reference to the presence of biases and the tendency of experienced VCs to rely on success patterns. It takes several years to gain relevant investing experience, create mental success models and identify potential patterns within them, but, once these cues are established, VCs face confirmation biases as they search for, interpret, favor and recall information that confirms or supports these patterns. This tendency may be exacerbated due to the interplay with availability biases because they can lead a venture capitalist to believe that a handful of examples, which are oftentimes the basis for their patterns, are more representative than they actually are. As a result, VCs might become closed-minded and less open to new concepts, which in turn may result in them overlooking novel, previously unseen opportunities. The last group of results concerns what can be summarized as education-specific performance dependencies. I find that there exists a positive correlation between a venture capitalist's level of education and their screening performance. While the screening performance of VCs without any degree is worst, it gradually increases with a bachelor's degree, master's or MBA degree and peaks with a Ph.D. degree. This might be due to the fact that the longer VCs spend in academia, the more they learn to

constantly challenge the status quo, develop new concepts and think them through. This explanation seems particularly true for Ph.D. graduates, who often spend close to a decade at universities, during which time they constantly train their brains to learn on an ongoing basis. In contrast, the brains of people who do not have the discipline required to spend years becoming an expert on a particular topic might be "lazy," and such individuals may instead rely on established patterns rather than spending the time to think something through on their own. Although the root cause might again be a reliance on patterns and established mental models, the reason for doing so might in this case not be based on overconfidence or availability biases but rather due to laziness to think through innovative and previously unseen concepts. Lastly, I find that STEM graduates perform better than business graduates or graduates with other degrees. This might be due to the fact that I only presented software companies, a field where VCs with an academic background in a related area might have an unfair advantage. Based on my survey results, the best-performing venture capitalist in terms of the generic screening of European early-stage software companies would have the following profile: an investment professional with a Ph.D. in STEM and less than 10 years' VC experience who works for an institutional VC firm with more than €500 million in AUM.

2.3.6.2 Limitations and avenues for future research

As with all empirical research, my work is subject to several limitations. I have consistently attempted to ensure robustness and validity of my study concerning the following four issues.

First, the XG algorithm and the best-performing investment professionals achieve the same AC and RE value of 80% in my study. While the resolution of 20% per increment is sufficient to conclude that the XG algorithm significantly outperforms the median and average investment professionals, this resolution is not granular enough to determine whether the algorithm outperforms even the best-performing VCs. As described previously, the resolution is due to the rather small company sample size and the success/failure ratio of 5:5 in my test

dataset. Therefore, I suggest either replicating my study with a larger company sample set while similarly ensuring a sufficiently large respondent sample or exploring fundamentally new ways of collecting VC success/failure predictions at scale that satisfy the benchmarking requirements identified above. Second, my study is focused on European early-stage software companies and is thus not representative for other geographies, stages or industries. Although I randomly tested my trained XG classifier with out-of-focus companies and still achieved similar (or occasionally slightly worse) results, I suggest retraining the algorithm with different datasets and benchmarking it with an equivalent group of investment professionals. I recommend changing only one of the three dimensions at a time to ensure comparability with my study. Third, I reduced the number of SC from all available data to a subset of information by following the substantial SC literature and conducting pre-tests with VCs. Moreover, at the end of the survey, I asked the respondents whether they would have required additional information to make an informed go/no-go screening decision and excluded those who answered "yes". Although scholars such as Zacharakis and Meyer (2000) and Åstebro and Elhedhli (2006) provide strong evidence that reduced availability of information does not negatively impact performance and I additionally ensured that perceived performance was not negatively impacted by omitted information, full certainty would require an A/B-test with A including all available SC and B including the reduced SC set. It might be the case that some VCs perform better with more information even though they do not perceive their improved performance. Finally, I used actual company information from between 2015 and 2016 and asked VCs to predict the outcomes of the anonymized companies as of January 2020. While I explicitly asked the respondents whether they had identified one or more of the companies and excluded them if the answer was "yes," I cannot be certain as to whether the remaining respondents either consciously identified them and did not indicate this or subconsciously identified them and did not notice. To circumvent this issue, I suggest collecting actual company information as of

today, collecting the VC predictions immediately after and then calculating the confusion matrix a few years later (or as soon as the success/failure outcomes can be determined).

In summary, my findings show that skepticism and fears of a potential performance drop due to the implementation of ML-based screening tools are unfounded. I hope that this study will contribute to the growing area of AI/ML in VC and that my empirical results will help to resolve the automation–control trade-off by creating the necessary trust. I am convinced that intelligent screening tools and automation more generally are crucial levers for scaling the traditional VC model and that their adoption is not an option but a necessity. People need to acknowledge that computers are superior when it comes to performing repetitive tasks and objectively processing large amounts of information, whereas humans are better in building relationships and understanding the nuances of an investment to ultimately make appropriate decisions and secure the best deals.

3 Conclusion

The limited availability of large-scale, high-quality private company and VC data has been a major constraint for researchers and practitioners alike. Existing data collection attempts are subject to trade-offs among sample size, level of detail and the freedom to share the resulting datasets. Moreover, only a limited number of datasets have been scrutinized with respect to their comprehensiveness and data quality. As a consequence, the wide range of frequently unverified datasets leads to different interpretations and conclusions. These conflicting interpretations subsequently result in a lack of trust in the underlying data and analyses thereof. Essay 1 sought to overcome data collection barriers and explored a replicable bottom-up data collection approach that results in a detailed, large-scale and freely sharable private company dataset. Essay 2 applied an established data verification approach to scrutinize the most prominent VC databases and determine their data quality. This approach is intended to assist researchers and practitioners to interpret available information more accurately. Given the assumption that the two preceding essays will contribute to creating the necessary trust in private company and VC datasets, Essay 3 showcased how ML algorithms leverage such information to scale the VC investment process. Essay 3 presents a comprehensive performance benchmarking between the best-performing ML algorithm and European VC investment professionals to promote further trust in data-driven approaches and accelerate their adoption.

3.1 Summary of research findings and contributions

The findings of this dissertation have several implications for academic researchers and VC practitioners. It provides three major building blocks that can help practitioners and academics to further gauge and unlock the potential of private company and VC data. Beyond collecting and verifying large-scale, high-detail and freely sharable datasets in Essays 1 and 2, I showcase how such information can be applied to empirically approach novel research

questions in Essay 1 and how VC practitioners can leverage ML algorithms on the basis of such datasets to improve their investment processes in Essay 3. I summarize the major findings and contributions of each essay in three separate paragraphs below.

Essay 1 explored a scalable approach to collecting private company data that overcomes common sample size limitations and makes it possible to resolve thus-far unanswered research questions. By explaining our bottom-up data collection approach in detail, we allow the reader to replicate it and thereby collect large private company datasets that can be freely shared. We applied the resulting dataset to determine "to what degree do findings on portfolio diversification translate from fVC to iVC?" and presented three valuable contributions as a byproduct of our innovative data collection method: First, we found that both investor types benefit from spreading their investments across industries and venture stages but that such an approach returns less value for iVCs than it does for fVCs. More experienced iVCs in our analysis performed better through less, comparatively, diversification of their investments, a signal that could be attributed to focusing on certain types of investments where they possess operational experience. Our results reflect the limited investment capacity of iVCs, which might push them to contribute to fewer rounds of investments despite the successful choice of ventures to invest. This likely prevents them from leveraging the full advantages of diversification. Second, our results indicate how the properties of the investment strategies employed by investors (iVCs in particular) shed more light on the exact benefits of diversification. For example, greater prior investment experience in terms of number of investments most likely benefits both types of investors by leading them to focus on fewer industries. In line with our previous findings, fVCs and iVCs are initially more successful when diversifying across industries; however, increasing investment experience reverses this effect and makes industry specialization increasingly more successful. To put it differently, prior investment experience acts as a substitute for the need to diversify across industries. Moreover, we find that iVCs, who

on average make investments in more mature ventures (i.e., ventures at *later* stages of their development) benefit more from a more diverse portfolio of such investments. Our results indicate that at later investment stages, it becomes increasingly difficult for iVCs to gain an industry- or stage-specific competitive advantage through specialization. A final feature that affects the impact of iVCs' portfolio diversification on performance is the average risk of the industries included in the portfolios of such investors. The higher the average risk of the industries included in a portfolio, the less successful diversification (eq. the more successful specialization) becomes. This might be explained with reference to the fact that the more complex and riskier the industries included in an investment portfolio are, the more effort it takes to fully understand the industry- and stage-specific challenges involved. Hence, investors may limit their "spread" across fewer industries or stages to ensure that they can devote sufficient attention to and have the capacity required to understand the ventures in which they have invested. Once an investor gains an understanding of a particular industry or stage, it likely provides a competitive advantage; thus, an iVC should focus future investment activities on similar settings (i.e., pursue industry and stage specialization rather than diversification). Third, we find that portfolio returns are, independent of investor types, highly skewed and distributed according to a power-law. They depend on one or a few home run investments per portfolio for both fVCs and iVCs. Despite similar return profiles, we observe that iVCs tend to minimize their downside risk rather than maximizing their upside potential, whereas fVCs focus on maximizing their upside potential without placing much emphasis on protecting their downside.

In Essay 2, we collected actual contracts and investment documentation from different VC partnerships and compared these documents with their characterisation in the eight most relevant VC databases. While the major driver of this study was showcasing a replicable data verification method that can be applied in a similar manner to any other dataset, our benchmarking results may help researchers and VC practitioners to better understand the coverage and quality of these frequently used databases and thus interpret the information

contained in them more accurately. More specifically, our results indicate that VS, PB and CB have the best coverage and are the most accurate databases across the dimensions of general company, founders and funding information. With respect to sampling biases, we found that greater financing rounds are more likely to be reported than lower ones. Similarly, financing round sizes and post-money valuations are more likely to be reported for greater financing rounds than for lower ones. Beyond these general patterns, we find a number of specific biases and sampling errors that should be considered when working with the databases under investigation. These findings served as a prerequisite for conducting the research presented in Essay 3, as previous studies have mainly focused on training ML algorithms for investment screening based on CB data without ensuring the comprehensiveness and considering the potential biases of this database. As a consequence of our results, we are aware of the shortcomings of the used databases and are able to purposefully complement or edit specific variables when putting together a training dataset.

On the basis of Essay 2, Essay 3 presented a comprehensive investment screening performance benchmarking between ML algorithms and human investment professionals. Initially, I conducted several interviews with VCs to understand the adoption of ML and data-driven screening approaches within their investment processes. The results clearly indicated that VCs are hesitant to adopt such novel tools, mainly due to a lack of trust in the underlying data quality and the absence of a comprehensive performance benchmarking study. Consequently, I assume that this study provides the missing building block required to create the necessary trust and accelerate the adoption of ML and data-driven screening tools. Although I found that my XG classification algorithm performed relatively 25% better than the *median* VC and 29% better than the *average* VC in terms of screening and selecting European early-stage software companies, by no means do I suggest replacing humans in the screening stage. Instead, I recommend an augmented approach where ML-based screening tools narrow the upper part of the deal funnel and present investment professionals with a selection of promising

opportunities. Using such an approach, new investment opportunities can be selected in an objective and highly efficient way, and, as a result, investment professionals can save substantial time, which can then be focused on appropriately evaluating a selection of high-potential deals. Moreover, they can use the freed-up resources to build stronger relationships with the identified entrepreneurial teams and put themselves in a better position to secure the most competitive deals. Instead of going broad and shallow by allocating limited resources to an ever-growing number of opportunities, the use of ML-based screening tools allows a venture capitalist to free up the time required to go narrow and deep on a selected number of opportunities while not overlooking promising deals. Ultimately, this data-driven approach can help VCs to scale their investment processes and eliminate potential biases in their screening. Similarly to Essay 1, this study yielded several valuable findings as byproducts. Namely, I provided a comprehensive characteristic-specific performance analysis with respect to VC investment professionals. My results show a) that institutional VCs perform better in the generic screening stage than CVCs; b) that after approximately a decade of VC experience, there exists a negative correlation between experience and screening performance; c) that there exists a positive correlation between a venture capitalist's highest level of education and their screening performance; and d) that STEM graduates perform better than business graduates or graduates with other degrees. Based on my benchmarking results, I find that the best-performing venture capitalist for the generic screening of European early-stage software companies would have the following profile: an investment professional with a Ph.D. in STEM and less than 10 years' VC experience who works for an institutional VC firm with more than €500 million in AUM.

3.2 Avenues for future research

This dissertation makes valuable contributions to the field of VC, but, naturally, it could not exhaust the many important issues addressed. While conducting the work for this dissertation, several avenues for future research emerged. I summarize these avenues in the

paragraphs below, each of which focuses on one of the three standalone essays presented in this dissertation.

Although the major motivation for Essay 1 was to overcome existing barriers to the collection of private data, the text focused on the application of the resulting dataset in order to render it suitable for submission to relevant academic journals. Consequently, the three avenues for future research described in Essay 1 are also related to the research question of portfolio diversification versus specialization for fVCs and iVCs rather than the data collection process itself. First, we assumed that our dependent variable *MaxFon* is a meaningful approximation of a VC's internal rate of return. We used this variable for lack of a better alternative; however, we believe that future efforts should be made to address the same research question with more accurate metrics. Second, there may exist endogeneity effects between the available capital per investor as well as the previous operational experience per investor and the pursued strategy. While the differences in terms of capital availability between iVCs and fVCs, among other reasons, may explain the distinct magnitudes of the diversification–performance relationships, a similar effect might exist within both investor groups that may in turn imply unobserved effects on investment strategy choices. Similarly, there may exist unobserved effects with respect to an investor's previous operational experience. In other words, the natural benefits of specialization might oppose the positive effects of diversification to a lesser extent for iVCs with higher capital availability (less operational experience) than for those with lower capital availability (more operational experience). Unfortunately, our data did not allow for the level of granularity required to fully address the issues of differences in capital availability and an investor's previous operational experience. Therefore, we suggest addressing the same research question with a more detailed dataset. Third, with respect to our bottom-up data collection approach and the comprehensiveness of the resulting dataset, we assumed that all VCs invest within the focal ecosystem of our study. Although our assumption is most likely true for the majority of iVCs, we likely overlooked a non-negligible proportion of fVC investments, as

many investors may be active across ecosystems. Still, the fact that our findings are in line with the previous literature on fVC confirms to an extent the validity of our approach. We hope that future research will explore alternative data collection approaches aimed at holistically representing all portfolios.

Essay 2 provided two major suggestions for future research: First, our original dataset consisted of 339 VC financing rounds in 108 companies and is thus not representative for all companies and financing rounds across Europe and certainly not globally. Although our study serves as a suitable approximation to better understand the coverage, quality and biases of the analyzed VC databases, we suggest replicating it with a larger, more diverse sample size. Second, we extended previous research by adding additional dimensions such as general company and founder-related information but are aware that there are even more variables to be considered and challenged. Therefore, we suggest that future research attempts to find ways to collect such original data and benchmarks the databases against them.

Essay 3 identified three potential avenues for future research: First, the trade-off between the number of companies included in the survey and the time required to screen these companies results in comparatively rough increments of 20% for AC and RE. To minimize these increments and achieve a more granular comparison, I suggest either replicating my study with a larger company sample set while similarly ensuring a sufficiently large respondent sample or exploring fundamentally new ways of collecting VC success/failure predictions at scale. Second, although I randomly tested my trained XG classifier with out-of-focus companies and still achieved similar (or occasionally slightly worse) results, I suggest retraining the algorithms with different datasets and benchmarking it with an equivalent group of investment professionals. I recommend only changing one of the three dimensions at a time to ensure comparability with my study. Lastly, I cannot be certain as to whether some survey respondents may have either consciously identified one or more of the 10 companies and did not indicate this or subconsciously identified them and did not notice. To mitigate this issue, I

suggest collecting actual company information as of today, collecting the VC responses immediately after and then calculating the confusion matrix in a few years later (or as soon as the success/failure outcomes can be determined).

Appendix A

Angellist. Established in 2010, the platform describes itself as "the world's largest startup community" which can be split into a talent and career page, an investing platform with information on startups and investors alike, and a product hunt page. Its database provides 84 variables on 4.9 million companies which it mainly sources through direct contributions from its community. Founders and investors can create their own profiles which are checked via some basic rule-based systems and by its team of approximately 60 employees.

CB-Insights. The provider describes itself as "a machine intelligence platform that catches every private company financing and angel investment". Its database launched in June 2009 and claims to cover "hundreds of thousands of companies". CB-Insights provides 104 variables per company which it mainly sources through web crawlers and manual desk research by its 250 employees. Besides its database, it provides market intelligence and a variety of reports.

Crunchbase. The provider describes itself as "the leading destination for company insights from early-stage startups to the Fortune 1000". Its database launched in July 2007 and claims to have more than 3.9 billion yearly updates and a coverage of more than 100,000 companies, more than 3,700 investors and more than 100,000 individual founders or managers. Per company, the database provides 112 different variables that are sourced via three approaches: 1) Community: Anyone can submit information. However, this information is subject to registration, social validation and are oftentimes reviewed by one of its employees. 2) Inhouse data team: A team of more than 150 employees manually collects information and inputs it onto the platform. 3) Machine learning models: It utilizes a range of web crawlers and deterministic as well as machine learning models in order to spot novel information, classify it and immediately fill it into the platform. Furthermore, these models help to validate data accuracy and alert its employees about anomalies and data conflicts.

Dealroom. Launched in 2013, Dealroom consolidates multiple data sources into one database which provides 189 variables per company on more than 460,000 ventures. It claims to have a particularly strong focus on Europe, which results from its community-driven approach. Founders and investors are incentivized to submit their data in order to have it considered in their prominent reports. Moreover, it enriches its database through automated crawlers which focus on social media, curated media and selected websites. It claims to leverage machine learning together with its team of approximately 15 employees, so as to maintain high data quality and to identify and correct potential issues.

Pitchbook. Launched in 2009, Pitchbook employs more than 700 people and provides data on global venture capital, private equity and public markets. It covers more than 2.2 million companies and provides up to 306 variables per company, which it collects via more than 650,000 web crawlers and a dedicated research team. As it is one of the major databases used by investors, Pitchbook directly validates the information via its investor relation teams. Furthermore, it has a quality assurance team that uses multiple validation methods and manual reviews to "vet every piece of data". Pitchbook claims to be a "one-stop-shop" for all VC related information, providing comprehensive information throughout the full venture life cycle, but also for fund-related metrics such as performance.

Preqin. The company provides financial data and information on the alternative assets market including fund, fund manager, investor and deal information in venture capital. Founded in 2003, Preqin provides information on more than 35,000 investors and up to 66 variables on "hundreds of thousands of companies". Preqin employs 500 people and collects data through dedicated research teams who curate the information in 1-to-1 conversations with market participants. While it also leverages web crawlers and machine learning models to automatically enrich its dataset, the company claims that every datapoint is manually checked before it reaches the platform. Anomaly detection models supposedly highlight potential issues and accelerate the investigation processes.

Tracxn. Founded in 2013, Tracxn describes itself as “a data-driven research platform that provides business updates and insights about startup companies”. It claims to track over 10 million companies globally and provides 104 variables per company. Tracxn sources its data mainly through a variety of web crawlers and machine learning models. It claims to be “the most automated data platform for private market investors”. Despite its high degree of automation, the company employs more than 900 people, more than 100 of them being domain experts who validate data and create research reports. It claims to focus on the earliest stages of companies and have them covered even before they receive their first financing round.

VentureSource. Founded in 1987, VentureSource is the oldest provider in our selection which, surprisingly, only employs 25 people. The company is owned by Dow Jones and describes itself as “the most accurate, comprehensive global database on companies backed by venture capital and private equity in every region, industry and stage of development.” It claims to have more than 67,000 companies and more than 20,000 investors in its dataset, and provides 154 variables per company. Please note that Dow Jones decided to discontinue the database as of March 31, 2020.

Literature

Achleitner, A.-K. (2001). Venture Capital. In *Handbuch Finanzierung* (pp. 513–529). Springer.

Achleitner, A., Braun, R., Behrens, J. H., and Lange, T. (2019). Enhancing innovation in Germany by strengthening the growth finance ecosystem. *Acatech Publication (Acatech STUDY)*.

Aghion, P., and Bolton, P. (1992). An incomplete contracts approach to financial contracting. *The Review of Economic Studies*, *59*(3), 473–494.

Alexy, O. T., Block, J. H., Sandner, P., and Ter Wal, A. L. J. (2012). Social capital of venture capitalists and start-up funding. *Small Business Economics*, *39*(4), 835–851.

Antretter, T., Sirén, C., Grichnik, D., and Wincent, J. (2018). *How individual business angels increase investment returns through angel networks: The impact of diversification and network centrality on portfolio performance*.

Arroyo, J., Corea, F., Jimenez-Diaz, G., and Recio-Garcia, J. A. (2019). Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. *Ieee Access*, *7*, 124233–124243.

Åstebro, T., and Elhedhli, S. (2006). The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, *52*(3), 395–409.

Atomico. (2019). State of European Tech 2019. Retrieved February 10, 2020, from https://2019.stateofeuropeantech.com/

Bachher, J. S., and Guild, P. D. (1996). Financing early stage technology based companies: investment criteria used by investors. *Frontiers of Entrepreneurship Research*, *996*.

Barnett, W. P., Greve, H. R., and Park, D. Y. (1994). An evolutionary model of organizational performance. *Strategic Management Journal*, *15*(S1), 11–28.

Bernile, G., Cumming, D., and Lyandres, E. (2007). The structure of private equity fund portfolios: Theory and international evidence. *Journal of Corporate Finance*, *4*, 564–590.

Beyer, D., Löwe, S., and Wendler, P. (2019). Reliable benchmarking: Requirements and solutions. *International Journal on Software Tools for Technology Transfer*, *21*(1), 1–29.

Bonini, S., and Capizzi, V. (2019). The role of venture capital in the emerging entrepreneurial finance ecosystem: future threats and opportunities. *Venture Capital*, *21*(2–3), 137–175.

Bonini, S., Capizzi, V., Valletta, M., and Zocchi, P. (2018). Angel network affiliation and business angels' investment practices. *Journal of Corporate Finance*, *50*, 592–608.

Boocock, G., and Woods, M. (1997). The evaluation criteria used by venture capitalists: evidence from a UK venture fund. *International Small Business Journal*, *16*(1), 36–57.

Braun, R., Jenkinson, T., and Stoff, I. (2017). How persistent is private equity performance? Evidence from deal-level data. *Journal of Financial Economics*, *123*(2), 273–291.

Braun, R., Weik, S., and Achleitner, A. (2019). Foreign Venture Capital in Europe: Consequences for Ventures' Exit Routes and Entrepreneurial Migration. *Available at SSRN 3415370*.

Brealey, R., Leland, H. E., and Pyle, D. H. (1977). Informational asymmetries, financial structure, and financial intermediation. *The Journal of Finance*, *32*(2), 371–387.

Breugst, N., Domurath, A., Patzelt, H., and Klaukien, A. (2012). Perceptions of entrepreneurial passion and employees' commitment to entrepreneurial ventures. *Entrepreneurship Theory and Practice*, *36*(1), 171–192.

Brush, C. G., and Vanderwerf, P. A. (1992). A comparison of methods and sources for obtaining estimates of new venture performance. *Journal of Business Venturing*, *7*(2), 157–170.

Bruvold, N. T., and Comer, J. M. (1988). A model for estimating the response rate to a mailed survey. *Journal of Business Research*, *16*(2), 101–116.

Buchner, A., Mohamed, A., and Schwienbacher, A. (2017). Diversification, risk, and returns in venture capital. *Journal of Business Venturing*, *32*(5), 519–535.

Bygrave, W. D. (1988). The structure of the investment networks of venture capital firms. *Journal of Business Venturing*, *3*(2), 137–157.

Catalini, C., Foster, C., and Nanda, R. (2018). *Machine intelligence vs. human judgement in new venture finance*. Mimeo.

Cook, C., Heath, F., and Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836.

Cressy, R., Malipiero, A., and Munari, F. (2014). Does VC fund diversification pay off? An empirical investigation of the effects of VC portfolio diversification on fund performance. *International Entrepreneurship and Management Journal*, *10*(1), 139–163.

Cressy, R., Munari, F., and Malipiero, A. (2007). Playing to their strengths? Evidence that specialization in the private equity industry confers competitive advantage. *Journal of Corporate Finance*, *13*(4), 647–669.

Croce, A., Guerini, M., and Ughetto, E. (2018). Angel Financing and the Performance of High-Tech Start-Ups. *Journal of Small Business Management*, *56*(2), 208–228.

Cumming, D., and Dai, N. (2011). Fund size, limited attention and valuation of venture capital backed firms. *Journal of Empirical Finance*, *18*(1), 2–15.

Cumming, D. J. (2006). The determinants of venture capital portfolio size: empirical evidence. *The Journal of Business*, *79*(3), 1083–1126.

Dalle, J.-M., Den Besten, M., and Menon, C. (2017). *Using Crunchbase for economic and managerial research*.

de Lange, D. E. (2019). A paradox of embedded agency: Sustainable investors boundary bridging to emerging fields. *Journal of Cleaner Production*, *226*, 50–63.

Dimov, D., and De Clercq, D. (2006). Venture capital investment strategy and portfolio failure rate: A longitudinal study. *Entrepreneurship Theory and Practice*, *30*(2), 207–223.

Dixon, R. (1991). Venture capitalists and the appraisal of investments. *Omega*, *19*(5), 333–344.

Domurath, A., and Patzelt, H. (2019). Founder non-international experience and venture internationalization. *Journal of International Entrepreneurship*, *17*(4), 494–519.

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R., and Kwan, I. (2002). Increasing response rates to postal questionnaires: systematic review. *Bmj*, *324*(7347), 1183.

Esparza, J., Scherer, S., Brechmann, A., and Schwenker, F. (2012). Automatic emotion classification vs. human perception: Comparing machine performance to the human benchmark. *2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA)*, 1253–1258. IEEE.

Franke, N., Gruber, M., Harhoff, D., and Henkel, J. (2006). What you are is what you like—similarity biases in venture capitalists' evaluations of start-up teams. *Journal of Business Venturing*, *21*(6), 802–826.

Freear, J., Sohl, J. E., and Wetzel Jr, W. E. (1994). Angels and non-angels: are there differences? *Journal of Business Venturing*, *9*(2), 109–123.

Fried, V. H., and Hisrich, R. D. (1994). Toward a model of venture capital investment decision making. *Financial Management*, 28–37.

Gastaud, C., Carniel, T., and Dalle, J.-M. (2019). The varying importance of extrinsic factors in the success of startup fundraising: competition at early-stage and networks at growth-stage. *ArXiv Preprint ArXiv:1906.03210*.

Ghassemi, M. M., Song, C., and Alhanai, T. (2020). *The Automated Venture Capitalist: Data and Methods to Predict the Fate of Startup Ventures*.

Gifford, S. (1997). Limited attention and the role of the venture capitalist. *Journal of Business Venturing*, *12*(6), 459–482.

Gompers, P., Kovner, A., and Lerner, J. (2009). Specialization and success: Evidence from venture capital. *Journal of Economics and Management Strategy*, *18*(3), 817–844.

Gompers, P., and Lerner, J. (2000). Money chasing deals? The impact of fund inflows on private equity valuation. *Journal of Financial Economics*, *55*(2), 281–325.

Gompers, P., and Lerner, J. (2001). The venture capital revolution. *Journal of Economic Perspectives*, *15*(2), 145–168.

Gompers, Paul A, Gornall, W., Kaplan, S. N., and Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, *135*(1), 169–190.

Gompers, Paul A, and Lerner, J. (1999). *What drives venture capital fundraising?* National bureau of economic research.

Gompers, Paul Alan, and Lerner, J. (2004). *The venture capital cycle*. MIT press.

Gorman, M., and Sahlman, W. A. (1989). What do venture capitalists do? *Journal of Business Venturing*, *4*(4), 231–248.

Hall, J., and Hofer, C. W. (1993). Venture capitalists' decision criteria in new venture evaluation. *Journal of Business Venturing*, *8*(1), 25–42.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Haunschild, P. R., and Sullivan, B. N. (2002). Learning from complexity: Effects of prior accidents and incidents on airlines' learning. *Administrative Science Quarterly*, *47*(4), 609–643.

Heberlein, T. A., and Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, 447–462.

Hellmann, T., Lindsey, L., and Puri, M. (2007). Building relationships early: Banks in venture capital. *The Review of Financial Studies*, *21*(2), 513–541.

Hisrich, R. D., and Jankowicz, A. D. (1990). Intuition in venture capital decisions: An exploratory study using a new technique. *Journal of Business Venturing*, *5*(1), 49–62.

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

Hsu, D. H. (2004). What do entrepreneurs pay for venture capital affiliation? *The Journal of Finance*, *59*(4), 1805–1844.

Humphery-Jenner, M. (2013). Diversification in private equity funds: On knowledge sharing, risk aversion, and limited attention. *Journal of Financial and Quantitative Analysis*, *48*(5), 1545–1572.

Hunter, D. S., Vielma, J. P., and Zaman, T. (2016). Picking winners using integer programming. *ArXiv Preprint ArXiv:1604.01455*.

Ingram, P., and Baum, J. A. C. (1997). Opportunity and constraint: organizations'learning from the operating and competitive experience of industries. *Strategic Management Journal*, *18*(S1), 75–98.

Jääskeläinen, M., Maula, M., and Seppä, T. (2006). Allocation of attention to portfolio companies and the performance of venture capital firms. *Entrepreneurship Theory and Practice*, *30*(2), 185–206.

Jackson III, W. E., Bates, T., and Bradford, W. D. (2012). Does venture capitalist activism improve investment performance? *Journal of Business Venturing*, *27*(3), 342–354.

Jennings, P., and Beaver, G. (1997). The performance and competitive advantage of small firms: a management perspective. *International Small Business Journal*, *15*(2), 63–75.

Jensen, M. C., and Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360.

Jiang, Y.-G., Ye, G., Chang, S.-F., Ellis, D., and Loui, A. C. (2011). Consumer video understanding: A benchmark database and an evaluation of human and machine performance. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 1–8.

Kanniainen, V., and Keuschnigg, C. (2003). The optimal portfolio of start-up firms in venture capital finance. *Journal of Corporate Finance*, *9*(5), 521–534.

Kanniainen, V., and Keuschnigg, C. (2004). Start-up investment with scarce venture capital support. *Journal of Banking and Finance*, *28*(8), 1935–1959.

Kaplan, S. N., and Lerner, J. (2016). *Venture capital data: Opportunities and challenges*. National Bureau of Economic Research.

Kaplan, S. N., and Schoar, A. (2005). Private equity performance: Returns, persistence, and capital flows. *The Journal of Finance*, *60*(4), 1791–1823.

Kaplan, S. N., Schoar, A., Gompers, P., Lerner, J., Brander, J. A., Amit, R., … Bygrave, W. D. (2005). On the positive role of financial intermediation in allocation of venture capital in a market with imperfect information. *Journal of Financial Economics*, *55*(4), 1935–1959.

Kaplan, S. N., Strömberg, P., and Sensoy, B. A. (2002). How well do venture capital databases reflect actual investments? *Available at SSRN 939073*.

Kirsch, D., Goldfarb, B., and Gera, A. (2009). Form or substance: the role of business plans in venture capital decision making. *Strategic Management Journal*, *30*(5), 487–515.

Knill, A. (2009). Should venture capitalists put all their eggs in one basket? Diversification versus pure-play strategies in Venture Capital. *Financial Management*, *38*(3), 441–486.

Korteweg, A., and Sorensen, M. (2010). Risk and return characteristics of venture capital-backed entrepreneurial companies. *The Review of Financial Studies*, *23*(10), 3738–3772.

Kortum, S., and Lerner, J. (2001). Does venture capital spur innovation? In *Entrepreneurial inputs and outcomes: New studies of entrepreneurship in the United States* (pp. 1–44). Emerald Group Publishing Limited.

Krishna, A., Agrawal, A., and Choudhary, A. (2016). Predicting the outcome of startups: less failure, more success. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 798–805. IEEE.

Lange, J. E., Bygrave, W., Nishimoto, S., Roedel, J., and Stock, W. (2001). Smart money? The impact of having top venture capital investors and underwriters backing a venture. *Venture Capital: An International Journal of Entrepreneurial Finance*, *3*(4), 309–326.

Lavender, J., Moore, C., Smith, K., and Eli, M. (2019). *Q4'19 Venture Pulse Report - Global Trends*. Retrieved from https://home.kpmg/xx/en/home/campaigns/2020/01/q4-venture-pulse-report-global.html

Lerner, J. (1995). Venture capitalists and the oversight of private firms. *The Journal of Finance*, *50*(1), 301–318.

Lussier, R. N. (1995). A nonfinancial business success versus failure prediction mo. *Journal of Small Business Management*, *33*(1), 8.

Maats, F., Metrick, A., Yasuda, A., Hinkes, B., and Vershovski, S. (2011). On the consistency and reliability of venture capital databases. *Unpublished Working Paper*.

MacMillan, I. C., Kulow, D. M., and Khoylian, R. (1989). Venture capitalists' involvement in their investments: Extent and performance. *Journal of Business Venturing*, *4*(1), 27–47.

MacMillan, I. C., Siegel, R., and Narasimha, P. N. S. (1985). Criteria used by venture capitalists to evaluate new venture proposals. *Journal of Business Venturing*, *1*(1), 119–128.

MacMillan, I. C., Zemann, L., and Subbanarasimha, P. N. (1987). Criteria distinguishing successful from unsuccessful ventures in the venture screening process. *Journal of Business Venturing*, *2*(2), 123–137.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Mason, C., and Harrison, R. (2017). Informal venture capital and the financing of emerging growth businesses. *The Blackwell Handbook of Entrepreneurship*, 221–239.

Mason, C. M. (2006). Informal sources of venture finance. In *The life cycle of entrepreneurial ventures* (pp. 259–299). Springer.

Mason, C. M., and Harrison, R. T. (2002). Is it worth it? The rates of return from informal venture capital investments. *Journal of Business Venturing*, *17*(3), 211–236.

Mason, W. S., Dressel, R. J., and Bain, R. K. (1961). An experimental study of factors affecting response to a mail survey of beginning teachers. *Public Opinion Quarterly*, 296–299.

Matusik, S. F., and Fitza, M. A. (2012). Diversification in the venture capital industry: leveraging knowledge under uncertainty. *Strategic Management Journal*, *33*(4), 407–426.

Muzyka, D., Birley, S., and Leleux, B. (1996). Trade-offs in the investment decisons of European venture capitalists. *Journal of Business Venturing*, *11*(4), 273–287.

Norton, E., and Tenenbaum, B. H. (1993). Specialization versus diversification as a venture capital investment strategy. *Journal of Business Venturing*, *8*(5), 431–442.

Paul, J. (2016). *4 Key Insights from Analyzing 5,000+ Cap Tables*.

Petty, J. S., and Gruber, M. (2011). "In pursuit of the real deal": A longitudinal study of VC decision making. *Journal of Business Venturing*, *26*(2), 172–188.

Pitchbook. (2019). *European Venture Report*. Retrieved from https://files.pitchbook.com/website/files/pdf/PitchBook_2019_Annual_European_Venture_Report.pdf

Pitchbook, and NVCA. (2019). *Venture Monitor Q4'2019*. Retrieved from https://files.pitchbook.com/website/files/pdf/Q4_2019_PitchBook_NVCA_Venture_Monitor.pdf

Pitchbook Research. (2020). Retrieved from https://pitchbook.com/blog/private-equity-vs-venture-capital-whats-the-difference

Plummer, J. L., and Walker, J. (1987). *QED report on venture capital financial analysis*. QED Research.

Poindexter, J. B. (1976). *The efficiency of financial markets: the venture capital case*. University Microfilms.

Prencipe, D. (2017). *The European venture capital landscape: an EIF perspective. Volume III: Liquidity events and returns of EIF-backed VC investments*. EIF Working Paper.

Rah, J., Jung, K., and Lee, J. (1994). Validation of the venture evaluation model in Korea. *Journal of Business Venturing*, *9*(6), 509–524.

Ramakrishnan, R. T. S., and Thakor, A. V. (1984). Information reliability and a theory of financial intermediation. *The Review of Economic Studies*, *51*(3), 415–432.

Retterath, A., and Braun, R. (2020). *Benchmarking Venture Capital Databases*.

Retterath, A., and Kavadias, S. (2020). How to Hit Home Runs: Portfolio Strategies and Returns in Formal and Informal Venture Capital. *Available at SSRN 3527412*.

Revilla, M., and Ochoa, C. (2017). Ideal and maximum length for a web survey. *International Journal of Market Research*, *59*(5), 557–565.

Ritter, J. (2015). To fly, to fall, to fly again. *The Economist*. Retrieved from https://www.economist.com/briefing/2015/07/25/to-fly-to-fall-to-fly-again

Rosenstein, J., Bruno, A. V, Bygrave, W. D., and Taylor, N. T. (1993). The CEO, venture capitalists, and the board. *Journal of Business Venturing*, *8*(2), 99–113.

Ross, S. (1976). The Arbitrage Theory of Capital Asset Pricing. *J. Econom. Theory*, *13*, 341–360.

Ruhnka, J. C., and Young, J. E. (1991). Some hypotheses about risk in venture capital investing. *Journal of Business Venturing*, *6*(2), 115–133.

Sahlman, W. A. (1990). The structure and governance of venture-capital organizations. *Journal of Financial Economics*, *27*(2), 473–521.

Sandberg, W. R., and Hofer, C. W. (1987). Improving new venture performance: The role of strategy, industry structure, and the entrepreneur. *Journal of Business Venturing*, *2*(1), 5–28.

Schmidt, C. M. (2019). *The impact of artificial intelligence on decision-making in Venture Capital Firms*.

Segal, G., Borgia, D., and Schoenfeld, J. (2005). The motivation to become an entrepreneur. *International Journal of Entrepreneurial Behavior and Research*.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science*, *9*(2), 277–293.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425–442.

Singer, E. (1978). Informed consent: Consequences for response rate and response quality in social surveys. *American Sociological Review*, 144–162.

Siskos, J., and Zopounidis, C. (1987). The evaluation criteria of the venture capital investment activity: An interactive assessment. *European Journal of Operational Research*, *31*(3), 304–313.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. *Australasian Joint Conference on Artificial Intelligence*, 1015–1021. Springer.

Sorenson, M. (2007). How smart is smart money? a two-sided matching model of venture capital. *Journal of Finance*, *62*(6), 272562.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. (2012). Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, *32*, 323–332.

Statista. (2019). Value of Venture Capital Investment in the US, 1995-2019.

Stuart, R., and Abetti, P. A. (1987). Start-up ventures: Towards the prediction of initial success. *Journal of Business Venturing*, *2*(3), 215–230.

Ter Wal, A. L. J., Alexy, O., Block, J., and Sandner, P. G. (2016). The best of both worlds: The benefits of open-specialized and closed-diverse syndication networks for new ventures' success. *Administrative Science Quarterly*, *61*(3), 393–432.

Thies, F., Huber, A., Bock, C., Benlian, A., and Kraus, S. (2019). Following the crowd—does crowdfunding affect venture capitalists' selection of entrepreneurial ventures? *Journal of Small Business Management*, *57*(4), 1378–1398.

Timmons, Jeffrey A, Muzyka, D. F., Stevenson, H. H., and Bygrave, W. D. (1987). Opportunity recognition: The core of entrepreneurship. *Frontiers of Entrepreneurship Research*, *7*(2), 109–123.

Timmons, Jeffry A, and Bygrave, W. D. (1986). Venture capital's role in financing innovation for economic growth. *Journal of Business Venturing*, *1*(2), 161–176.

Tyebjee, T. T., and Bruno, A. V. (1984). A model of venture capitalist investment activity. *Management Science*, *30*(9), 1051–1066.

Walston, J. T., Lissitz, R. W., and Rudner, L. M. (2006). The influence of web-based questionnaire presentation variations on survey cooperation and perceptions of survey quality. *Journal of Official Statistics*, *22*(2), 271.

Wells, W. A. (1974). *Venture capital decision-making*. Carnegie-Mellon University.

Wilson, K. E., and Silva, F. (2013). *Policies for Seed and Early Stage Finance*.

Winkler, H.-J., Rieger, V., and Engelen, A. (2019). Does the CMO's personality matter for web traffic? Evidence from technology-based new ventures. *Journal of the Academy of Marketing Science*, 1–23.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education.

Wright Robbie Ken, M. (1998). Venture capital and private equity: A review and synthesis. *Journal of Business Finance and Accounting*, *25*(5-6), 521–570.

Yammarino, F. J., Skinner, S. J., and Childers, T. L. (1991). Understanding mail survey response behavior a meta-analysis. *Public Opinion Quarterly*, *55*(4), 613–639.

Yang, Y., Narayanan, V. K., and De Carolis, D. M. (2014). The relationship between portfolio diversification and firm value: The evidence from corporate venture capital activity. *Strategic Management Journal*, *35*(13), 1993–2011.

Zacharakis, A. L., and Meyer, G. D. (1998). A lack of insight: do venture capitalists really understand their own decision process? *Journal of Business Venturing*, *13*(1), 57–76.

Zacharakis, A. L., and Meyer, G. D. (2000). The potential of actuarial decision models: can they improve the venture capital investment decision? *Journal of Business Venturing*, *15*(4), 323–346.

Zacharakis, A. L., and Shepherd, D. A. (2001). The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, *16*(4), 311–332.