CHAPTER 1

Literature Review

In this chapter, we first review the startup investment literature to develop criteria to evaluate our Venture Capital (VC) investment screening system. We then turn our focus to determining the best techniques to use to create this system, which we break down into three intercorrelated areas: feature selection, data sources and classification algorithms.

1. Criteria Selection. VC firms review many potential investment candidates to shortlist for investment. Traditional screening methods involve referral, networking and Internet search. These screening methods are highly time-consuming and subject to human selection biases. Based on our review, we believe that a superior system can be produced and should be assessed on the basis of its efficiency, robustness, and predictive power.

2. Feature Selection. VC is a key driver of startup development but our understanding of factors that influence VC firms' investment decisions and the subsequent performance of those investments is incomplete. Based on our review of the literature, we propose a hierarchical framework that includes a variety of features that have been previously indicated to be relevant to this investment screening problem. At a high-level, our framework incorporates determinants of startup potential and signals that influence investment confidence.

3. Data Sources. Startup performance is a multi-faceted problem and different data sources provide insights into different actors, relationships and attributes. Our review focuses on novel online data sources which have the potential to transform entrepreneurship and VC research. Preliminary evidence suggests that the online startup databases CrunchBase and AngelList are promising and likely to provide a comprehensive feature set that can form the basis of our system. Other sources like PatentsView, Twitter, LinkedIn, and PrivCo are considered.

4. Classification Algorithms. Predicting startup performance is a difficult problem for humans. After all, a high percentage of even VC-backed startups still fail. However, machine learning techniques have been recently used in other areas of finance (e.g. in the public markets) with some success. We cross-reference the characteristics of our intended dataset with the characteristics of common supervised classification algorithms. Our analyses suggest that we should expect Random Forests, Support Vector Machines and Artificial Neural Networks to be most suitable for our system.

1.1 Criteria Selection

Venture Capital (VC) financing has lagged behind other forms of high finance (e.g. bond trading, loan applications, insurance) in adopting computational analytics to aid decision-making. Banks are now able to evaluate personal loan requests in minutes while VC firms take far longer to put together deals, sometimes months. While these are markedly different forms of finance (VC has a longer return period, larger investments, higher risk profiles), a more data-informed and analytical approach to venture finance is still foreseeable.

In this section, we provide an introduction into VC firm strategy and review the existing state of the VC investment process. We find that analytical tools are nascent and use of analytics in industry is limited. To date only a small handful of VC firms have publicly declared their use of computational analytical methods in their decision making and investment selection process. We explore why the use of data mining in the VC industry is limited and we develop criteria by which we can judge a VC investment screening system to be successful.

1.1.1 Venture Capital Industry

Early-stage investment is a key driving force of technological innovation and is vitally important to the wider economy, especially in high-growth and technology intensive industries (e.g software, medical and agricultural technologies). VC is a form of private equity, a medium to long-term form of finance provided in return for an equity stake in potentially high growth companies [26]. Reported US VC investments in 2015 totalled US$60 billion [26].

Figure 1.1 illustrates the typical structure of a VC Fund. A typical fund is managed by a VC Firm (legally referred to as a General Partner) consisting of several investment partners. The fund itself (the Limited Partnership) is essentially an investment fund raised from various institutional investors such as

pension funds, university endowments and family offices (legally referred to as Limited Partners). Beyond fundraising the main responsibilities of investment partners (also referred to as General Partners) are sourcing investment opportunities, making investment decisions and taking board membership to assist the management of investee private companies (also referred to as Portfolio Companies).

Figure 1.1

1.1.2 Venture Capital Firm Strategy

Typically, VC firms are reliant on a small number of high-risk investments to produce outsized returns through successful exit events. A common rule-of-thumb is that given a portfolio of ten startup companies: three will fail entirely, three will remain active but will not be very profitable, three will be active and profitable, and one highly successful startup will provide the investor with a multiple return on all of the investments [**stone2014**]. In comparison to other traditional investment classes, VC financing is heavily biased towards control at the expense of risk mitigation. Although VC firms tend not to take majority stakes in startups, they exert their influence through significant minority stakes, board membership, their relative seniority to the companys founders, and through leveraging their business networks [14].

Despite VC firms, often significant, influence on the trajectory of their investments, they are still highly selective of the companies that they invest in. Although rarely reported, a small number of studies show VC investment rates vary between 1.5-3.5% of proposals considered [**stone2014**]. Accordingly, traditional venture finance is a very labour intensive and time consuming process involving extensive due diligence on behalf of the investor [**fried1994**]. The VC investment process involves several main stages: deal origination, screening, evaluation, structuring (e.g., valuation, term sheets), and post investment activities (e.g., recruiting, financing).

1.1.3 Current Venture Capital Systems

Early-stage investment is characterised by a large number of investment candidates, high degree of uncertainty; a lack of reliable data on company performance (particularly financial performance); and a high time-cost of undertaking due diligence. This makes for a complicated origination and screening process. While

referral from trusted sources (e.g., entrepreneurs, accountants, lawyers, other investors) is often used to screen opportunities, as the cost of starting businesses dramatically decreases investors are faced with an increasingly large number potential businesses and investment opportunities to assess and evaluate. Such a proliferation has led to an information overload problem in venture finance.

Despite evidence that VC firms could benefit from increased use of data mining, it appears few are interested in advanced data analytics. Stone [**stone2014**] interviewed Fred Wilson of Union Square Ventures who said: "We have not been able to quantify [startup potential]. We havent even tried. Although I am sure someone could do it and they might be very successful with it. To us, the ideal founding team is one supremely talented product oriented founder and one, two, or three strong developers, and nothing else." Likewise, when asked, Chris Dixon of Andreessen Horowitz said: "Ive seen a few attempts to do it quantitatively but I think those are often flawed because the quantitatively measurable things are either obvious, irrelevant, or suffer from overfitting (finding patterns in the past that dont carry forward in the future".

Similarly, while recently new software tools have been developed to assist VC firm, there is limited evidence of their adoption.

1.1.4 Proposed Criteria

Based on our review of the VC industry and current VC origination and screening processes, we have developed criteria on which we can evaluate our proposed system.

1. Efficiency. Our system must be more efficient than traditional, manual investment screening by referral and technology scan (e.g. Google search, media, databases). This means that it needs to be able to provide enough information – both observations and features – to be able to meet similar levels of accuracy.

2. Robustness. Our system must be robust enough to be reliable over time. The system must provide a generalised, robust solution for investors that does not require significant technical knowledge to use, and is not overfitted to a specific time-period or data source.

3. Predictive Power. Our system must be consistently accurate at identifying a variety of high-potential investment candidates. The system should be robust to different forecast windows (i.e. exit in three years from now)

as VC firms make investment decisions with different periods so they can strategically manage the investment horizons of their funds.

1.2 Feature Selection

Our understanding of the factors that influence Venture Capital (VC) investment decisions and the subsequent performance of those investments is incomplete. We believe a diverse range of features is critical to developing accurate models of startup performance and investment decisions.

Prior work focuses on basic company features (e.g. the headquarters' location, the age of the company) for startup investment predictive models [6, 16]. Semantic text features (e.g. patents, media) [19, 44] and social network features (e.g. co-investment networks) [41, 11, 43] may also predict startup investment. We expect a model that includes semantic text and social network features alongside basic company features could lead to better startup investment prediction.

We propose a conceptual framework that builds upon previous work to ensure that we include a comprehensive and relevant set of features in our investment screening system. Ahlers et al. [1] developed a conceptual framework for funding success on equity crowdfunding platforms. Their framework has two factors: venture quality and level of uncertainty. The first factor is based on work by Baum and Silverman [5] that suggests key determinants of startup potential are human capital, alliance (social) capital, and intellectual (structural) capital. The second factor is based on investors' confidence in their estimation of startup potential.

We seek to generalise Ahlers' framework [1] beyond equity crowdfunding. While the first factor of Ahlers' framework (venture quality) applies to startups of all stages, Ahlers operationalise their second factor with respect to whether startups offer an equity share in their crowdfunding, and whether they provide financial projections. These features are specific to equity crowdfunding. We propose an extension of Ahlers' framework that generalises and develops this second factor. We describe investment confidence as a product of third party validation, historical performance and contextual cues. Our proposed framework is depicted in Figure **??**.

Next, we must operationalise this conceptual framework into features that we can incorporate into our machine learning model. Table 1.1 shows a review of features tested in previous studies of startup investment. In Appendix **??**, we describe each of these features and outline theoretical and empirical evidence that justify their inclusion in our conceptual framework.

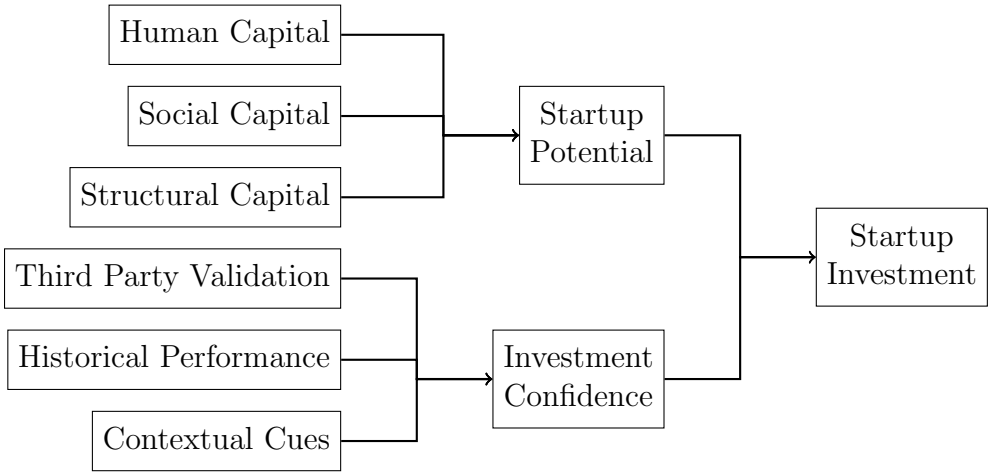

Figure 1.2: Proposed conceptual framework for startup investment. We adapt the framework proposed by Ahlers et al. [1], originally based on work by Baum and Silverman [5]. For an extended version of this framework, please refer to Figure **??**.

Feature selection is critical to the success of our proposed conceptual framework. In this section, we have built on the framework proposed by Ahlers et al. [1] in several ways. First, our framework generalises the "Investment Confidence" factor for startups seeking any type of investment (not just equity crowdfunding). Second, our framework has greater depth. Where Ahlers uses one or two features for each factor in their model (e.g. "% Nonexecutive board" represents "Social (alliance) capital"), we perform a review of many features employed in this area and perform a higher degree of classification. For example, in our proposed framework "Social (alliance) capital" is composed of "Social influence" and "Strategic alliances", each of which will also be composed of several features (e.g. "Twitter followers", "Average Tweets per day").

1.3 Data Sources

Predicting startup investment and performance is a complex and difficult task. There are many features that can influence startup investment decisions. Capturing the diversity of these features is critical to developing accurate models. Accordingly, this task will likely involve data collection from multiple data sources. Appropriate selection of these data sources is important because different data sources provide insights into different actors, relationships and attributes.

Previous studies in this field have been limited by data sources restricted in

Features	Results from Studies	
	Significant	Non-Significant
Startup Potential		
Human Capital		
Founders' Capabilities	[6, 3, 16]	[34, 12]
NED Capabilities	[5]	[1, 3]
Staff Capabilities	[6, 3, 12]	[1]
Social Capital		
Strategic Alliances	[5]	-
Social Influence	[6, 3, 11, 43]	-
Structural Capital		
Patent Filings	[19, 20, 5]	[1, 16]
Investment Confidence		
Third Party Validation		
Investment Record	[1, 6, 13, 19, 12]	-
Investor Reputation	[3, 41, 20]	[19]
Media Coverage	[6]	[3]
Awards and Grants	[1]	-
Historical Performance		
Financial Performance	[6, 5]	-
Non-Financial Performance	[3, 16]	[19]
Contextual Cues		
Competitor Performance	[34, 13, 16]	[6, 12]
Broader Economy	[6, 13, 19, 12, 20]	[34, 1]
Local Economy	[34, 6, 13, 16, 19]	-

Table 1.1: Features relevant to startup investment. We review thirteen empirical studies that investigate drivers of startup investment. For each study, we note whether included features have a significant effect on the startup investment model. We classify identified features according to our proposed conceptual framework.

sample size. Most studies have samples of fewer than 500 startups [1, 16], or between 500 and 2,000 startups [19, 43, 3, 41, 13], and only a few have large scale samples (more than 100,000 startups) [34, 11]. Sample size is more critical to model development than the sophistication of machine learning algorithms or feature selection [9]. Startups databases (e.g. CrunchBase) and social networks (e.g. Twitter) offer larger data sets than those previously studied. We expect data collected from these sources will lead to the discovery of additional features

and higher accuracy in startup investment prediction.

In Table 1.2, we outline the characteristics of relevant data sources and how they could contribute to our chosen features. In this section, we describe desirable characteristics of data sources for this task, review potentially relevant data sources, and ultimately determine which data sources are most likely to suit the characteristics of this task.

1.3.1 Source Characteristics

Entrepreneurship research is transforming with the availability of online data sources: databases, websites and social networks. Entrepreneurship studies have historically relied on surveys and interviews for data collection. Measures of human capital (e.g. founders' capabilities), strategic alliances, and financial performance are difficult to capture elsewhere. However, the trade-off for access to these features is that surveys and interviews are time-consuming and costly to implement. While online surveys address some of these issues, it is still difficult to motivate potential participants to contribute. Online data sources like startup databases and social networks are efficient because collecting data is a secondary function of users interacting with these sources. Researchers can also collect data from these sources automatically and at scale. For these reasons, we only consider online data sources for inclusion in this study, specifically crowd-sourced startup databases (e.g. CrunchBase, AngelList), social networks (e.g. Twitter, LinkedIn), government patent databases (e.g. PatentsView) and private company intelligence providers (e.g. PrivCo). In Appendix we review the characteristics of each of these data sources commonly used in entrepreneurship research.

1.3.2 Source Evaluation

Entrepreneurship and Venture Capital (VC) research is primed to take advantage of the availability of new online data sources. We evaluated relevant data sources for their suitability to predicting startup investment. Startup databases CrunchBase and AngelList provide the most comprehensive set of features. There are small differences between the features recorded by each. CrunchBase has slightly more coverage and tracks media better but lacks AngelList's social network. At least one startup database should be used and either are satisfactory. Of the other data sources we review, PatentsView is the most promising. PatentsView provides comprehensive patent information, though it could prove difficult matching identities to other sources. Other data sources are less promising because of access issues. LinkedIn cannot be easily collected now the API is deprecated.

Properties	Startup Databases		Social Media		Other Sources	
	CrunchBase	AngelList	LinkedIn	Twitter	PatentsView	PrivCo
Features						
Startup Potential						
Human Capital						
Founders' Capabilities	✓	✓	✓✓	✗	✗	✗
NED Capabilities	✓	✓	✓✓	✗	✗	✗
Staff Capabilities	✓	✓	✓✓	✗	✗	✗
Social Capital						
Social Influence	✓	✓✓	✓✓	✓✓	✗	✗
Strategic Alliances	✓	✓	✗	✗	✓	✗
Structural Capital						
Patent Filings	✗	✗	✗	✗	✓✓	✗
Investment Confidence						
Third Party Validation						
Investment Record	✓✓	✓✓	✗	✗	✗	✓
Investor Reputation	✓	✓✓	✓	✗	✗	✗
Media Coverage	✓✓	✓	✗	✓	✗	✗
Awards and Grants	✓	✗	✗	✗	✗	✗
Historical Performance						
Financial Performance	✗	✗	✗	✗	✗	✓✓
Non-Financial Performance	✓✓	✓✓	✓	✗	✗	✓
Contextual Cues						
Competitor Performance	✓	✓	✗	✗	✗	✗
Broader Economy	✓	✓	✗	✗	✗	✗
Local Economy	✓	✓	✗	✗	✗	✗
Ease of Use						
Cost Effective	✓	✓✓	✓	✗	✓✓	✗
Time Efficient	✓✓	✓✓	✗	✓✓	✓✓	✗
Accurate Data	✓	✓	✓✓	✓✓	✓✓	✓✓
Large Data Set	✓✓	✓✓	✓✓	✓✓	✓✓	✓

Table 1.2: Data sources relevant to startup investment. We review six data sources commonly used in entrepreneurship research for their suitability for our startup investment task. We evaluate data sources for their ability to provide relevant features for our analyses and for their ease of use in data collection. We exclude offline sources from our analyses. Ratings are: ✗ = poor, ✓ = satisfactory, ✓✓ = good.

Twitter provides social network topology and basic profile information through its free API but does not provide access to historical tweets. Financial reports are too expensive for the purposes of this study.

1.4 Classification Algorithms

Predicting startup performance is a difficult problem for humans. Computational analytics have been heavily deployed in high finance and we believe there is scope for applying related techniques to improve upon investment decision making in the domain of venture finance. Machine learning is characterised by algorithms that improve their ability to reason about a given phenomenon given greater observation and/or interaction with said phenomenon. Mitchell provides a formal definition of machine learning in operational terms: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [25].

Machine learning algorithms can be classified based on the nature of the feedback available to them: supervised learning, where the algorithm is given example inputs and desired outputs; unsupervised learning, where no labels are provided and the algorithm must find structure in its input; and reinforcement learning, where the algorithm interacts with a dynamic environment to perform a certain goal. These algorithms can be further categorised by desired output: classification, supervised learning that divides inputs into two or more classes; regression, supervised learning that maps inputs to a continuous output space; and clustering, unsupervised learning that divides inputs into two or more classes.

We evaluated common machine learning algorithms with respect to their suitability for predicting startup investment. In Table 1.3, we rank these algorithms by cross-referencing their assumptions and properties with the task characteristics. In the following sections, we describe the characteristics of the startup investment prediction task, review common machine learning algorithms, and determine which algorithms are most likely to suit the characteristics of this task.

1.4.1 Task Characteristics

Machine learning tasks are diverse. Our investigation into startup investment is a task that suits supervised machine learning algorithms. We will manipulate the data we collect into a single labelled data set. Startups will be labelled

Criteria	Machine Learning Algorithms													
	NB		LR		KNN		DT		RF		SVM		ANN	
Data Set Properties	**2**		4		6		**2**		**1**		5		7	
Missing Values	✓✓	[22]	✓	-	✗	[22]	✓✓	[22]	✓✓	[38]	✓	[22]	✗	[22]
Mixed Feature Types	✓✓	[22]	✓✓	-	✓✓	[22]	✓✓	[22]	✓	[38]	✓	[22]	✓	[22]
Irrelevant Features	✗	[22]	✗	[23]	✓	[22]	✓✓	[22]	✓✓	[38]	✗	[22]	✗	[22]
Imbalanced Classes	✓✓	-	✓✓	-	✗	-	✗	[22]	✓	[38]	✓✓	[22]	✓	[22]
Algorithm Properties	**2**		**1**		5		5		**2**		5		**2**	
Predictive Power	✗	[9]	✓	[9]	✓	[9]	✗	[22]	✓✓	[9]	✓✓	[9]	✗✓	[9]
Interpretability	✓✓	[22]	✓✓	[23]	✗	[22]	✓✓	[22]	✓	[23]	✗	[22]	✗	[22]
Incremental Learning	✓✓	[22]	✓✓	-	✓✓	[22]	✓	[22]	✓	-	✓	[22]	✓✓	[22]
Overall	**2**		**2**		6		4		**1**		5		6	

Table 1.3: Evaluation of machine learning algorithms for startup investment prediction. We review seven common supervised machine learning algorithms for their suitability for our startup investment task. We evaluate algorithms for their robustness to the structure of the data set and their appropriateness for the constraints of our implementation. We rank the algorithms according to the sum of these measures (in each section and overall) and bold highly-ranked algorithms. Ratings are: ✗ = poor, ✓ = satisfactory, ✓✓ = good. Algorithms are: NB = Naive Bayes, LR = Logistic Regression, KNN = K-Nearest Neighbours, DT = Decision Trees, RF = Random Forests, SVM = Support Vector Machines, ANN = Artificial Neural Networks.

based on whether they are acquired or have had an IPO at a later time. The key objective of machine learning algorithm selection is to find algorithms that make assumptions consistent with the structure of the problem (e.g. tolerance to missing values, mixed feature types, imbalanced classes) and suit the constraints of the desired solution (e.g. time available, incremental learning, interpretability). In the following sections, we outline the characteristics of supervised learning tasks relevant to our startup investment prediction task.

1.4.1.1 Data Set Properties

While data sets can be pre-processed to assist with their standardisation, some types of data sets are still better addressed by particular algorithms. Data set properties like missing data, irrelevant features, and imbalanced classes all have an effect on classification algorithms. Data sets often have missing values, where no data is stored for a feature of an observation. Missing data can occur because of non-response or due to errors in data collection or processing. Missing data has different effects depending on its distribution through the data set. Public data sets, like startup databases and social networks, are typically sparse with missing entries despite their scale. Therefore, robustness to missing values is a desirable property of our algorithm. Despite efforts to only include features that have theoretical relevance, machine learning tasks often include irrelevant features. Irrelevant features have no underlying relationship with classification. Depending on how they are handled they may affect classification or slow the algorithm. We expect irrelevant and non-orthogonal features in our data set because our proposed framework includes features that have not been thoroughly tested in the literature. Therefore, robustness to irrelevant features is a desirable property of our algorithm. Data sets are not usually restricted to containing equal proportions of different classes. Significantly imbalanced classes are problematic for some classifiers. In the worst case, a learning algorithm could simply classify every example as the majority class. Our data set is not dramatically imbalanced overall, but when looking at funding status for different funding rounds it is significantly imbalanced. Therefore, robustness to imbalanced classes is a desirable property of our algorithm.

1.4.1.2 Algorithm Properties

The desired properties of machine learning algorithms are related to the business problems that are being addressed. Predictive power, interpretability and processing speed are all desirable characteristics but involve trade-offs and must be prioritised. Predictive power is the ability of a machine learning algorithm to

correctly classify new observations. Predictive power can be evaluated in many ways. As our data set is likely to have an imbalanced class distribution, we will evaluate predictive power based on balanced metrics like Area under the Receiver-Operator Curve and the F1 Score. If a model has no predictive power, the model is not representing the underlying process being studied. For this reason, predictive power is a desirable property of our algorithm. However, if multiple algorithms provide similar predictive power other selection criteria become significant. Interpretability is the extent to which the reasoning of a model can be communicated to the end-user. There is a trade-off between model complexity and model interpretability. Some models are a "black box" in the sense that data comes in and out but the model cannot be interpreted. For this study, it is a key objective that we improve our understanding of the determinants of startup investment. Therefore, interpretability is a desirable property of our algorithm. Finally, processing speed is another desirable property, especially when handling real-time data or when there is a need to run exploratory analyses on the fly. In this case, processing speed is not critical because generally Venture Capital (VC) investment decisions are made over weeks and months, though there is some need for the data set to be updated with new information as it becomes available.

1.4.2 Algorithm Characteristics

Supervised machine learning are algorithms that reason about observations to produce general hypotheses that can be used to make predictions about future observations. Supervised machine learning algorithms are diverse, from symbolic (Decision Trees, Random Forests) to statistical (Logistic Regression, Naive Bayes, Support Vector Machines), instance-based (K-Nearest Neighbours), and perceptron-based (Artificial Neural Networks). In Appendix C, we describe each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in applications relevant to startup investment.

1.4.3 Algorithm Evaluation

We evaluated supervised learning algorithms for their suitability in startup investment prediction. While our evaluation gives us directionality of fit, we hesitate to discard algorithms based on our literature review. Algorithm selection is complex and preliminary testing will provide clarity as to which algorithms should be used. In addition, larger training sets and good feature design tend to outweigh

algorithm selection [9]. With those concessions in mind, our findings suggest we expect Random Forests, Support Vector Machines and Artificial Neural Networks to produce the highest classification accuracies. An ensemble of these algorithms may improve accuracy further, though at the cost of computational speed and interpretability. We may expect Random Forests to outperform the other two algorithms due to robustness to missing values and irrelevant features and native handling of discrete and categorical data. However, Random Forests are not highly interpretable so Decision Trees and Logistic Regression may be preferable for exploratory analysis of the data set.

1.5 Research Gap

The Venture Capital (VC) industry requires better systems and processes to efficiently manage labour-intensive tasks like investment screening. Existing approaches in the literature to predict startup performance have three common limitations: small sample size, a focus on very early stage investment, and incomplete use of features. In addition, there is little evidence that previous research has been translated into systems that are able to assist investors directly. We conducted a literature review to determine how to address these limitations and produce a system that will assist VC firms in originating and screening investment candidates.

Firstly, we reviewed the business problem and developed three criteria that will help us evaluate our system: efficiency, robustness and predictive power. Secondly, we developed a conceptual framework of predicting startup performance that incorporates determinants of startup potential and signals that influence investment confidence. This framework informs our feature selection. We then assessed potential data sources and found preliminary evidence that suggests that the startup databases CrunchBase and AngelList are promising and likely to provide a comprehensive feature set that can form the basis of our system. Finally, we reviewed supervised machine learning techniques applied to startup investment and other areas of finance. Our analyses suggested that we should expect Random Forests, Support Vector Machines and Artificial Neural Networks to be most suitable for our system.

Based on this literature review, we believe it is now possible to address previous limitations in this domain and produce an investment screening system that is efficient, robust and powerful. In the next chapter, we will outline the process by which we attempt to develop that system.

APPENDIX A

Feature Selection

We develop a conceptual framework relating startup potential and investor confidence to startup investment. We will operationalise this conceptual framework into features that can be incorporated into our machine learning model. To do this, we review features that have been tested in previous studies related to startup investment or performance. In the following sections, we describe each of these features and outline conceptual and empirical evidence that justify their inclusion in our conceptual framework. Figure A.1 depicts how these features can be incoprorated into our conceptual framework.

A.1 Venture Quality

A.1.1 Human Capital

Human capital is critical to early-stage startups that have limited resources and are changing constantly. Startups are composed of founders, non-executive directors (NED) that may be investors or advisers, and staff. Each of these parties makes a contribution to the human capital of the startup. The human capital of these parties can generally be categorised three ways: education, prior experience, and synergies as a team.

Founders' Capabilities Founders play multiple roles in early-stage startups, driving many aspects of the business growth and development. Accordingly, the human capital of founders has been shown to affect startup investment success. In particular, education of founders is a key signal. The number of degrees attained by founders is predictive of success [6, 16], as is whether a founder has obtained an MBA [6]. In addition, past entrepreneurial experience seems to be a predictive factor [16] though there is some evidence to dispute this [34]. Finally, the number of founders seems to be correlated

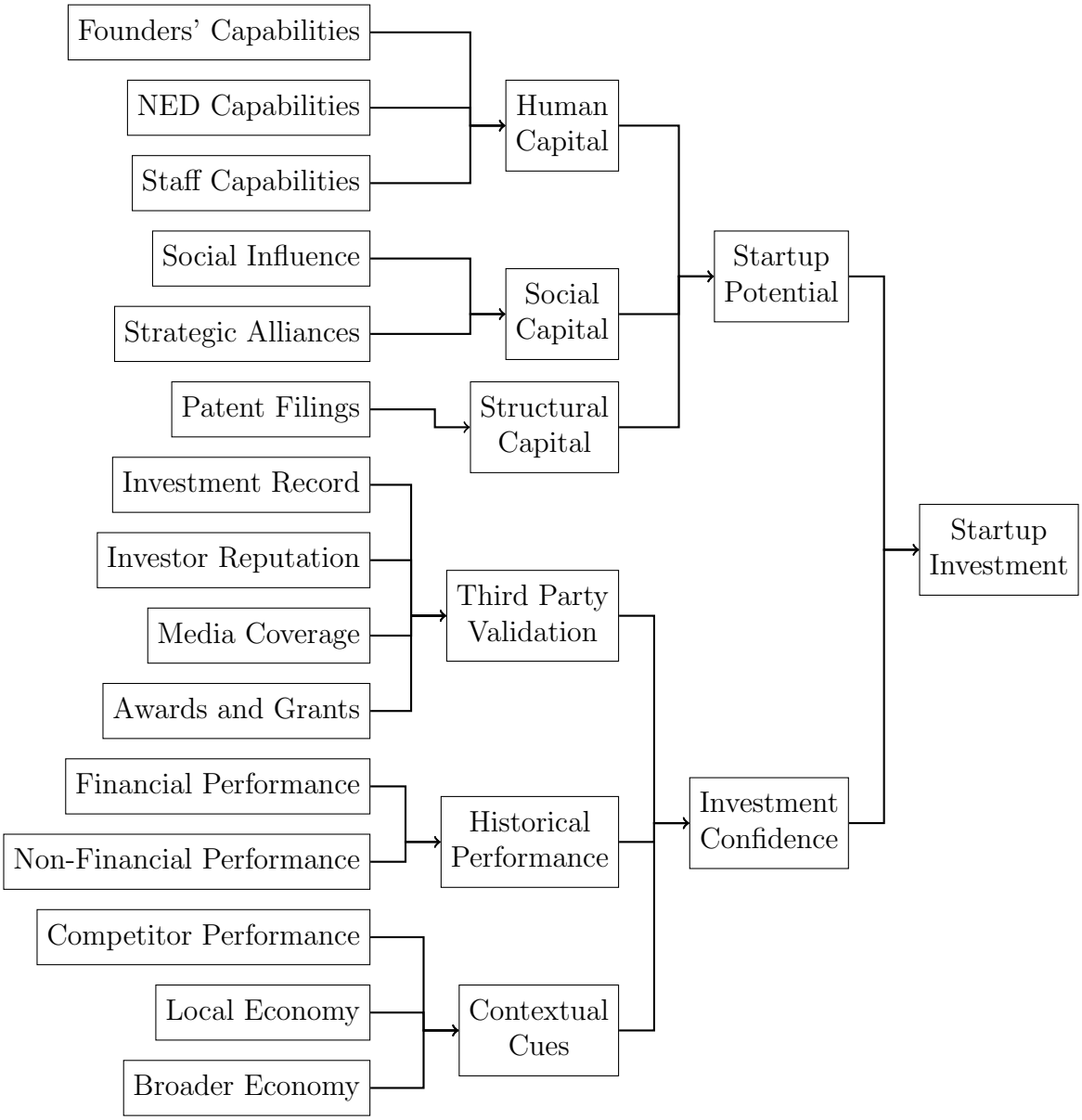

Figure A.1: Proposed conceptual framework for startup investment. This extended version of the framework includes features identified by empirical studies of startup investment. We adapt the framework proposed by Ahlers et al. [1], originally based on work by Baum and Silverman [5].

to startup success [6], though the underlying relationship may be more nuanced, and could be related to the distribution of team skillset.

Non-Executive Directors' Capabilities The boards of startups are smaller and have a higher concentration of ownership than those of well-established companies [21]. Startups lack corporate skills such as finance, human resources, information technology and legal expertise. Especially if founders are relatively inexperienced, they may look to the board to provide these skills. As a result, there is more overlap between governance and operational roles and directors may have greater influence on company performance through greater involvement in decision making [21]. Startups with more experienced directors are more successful at raising funds [5].

Staff Capabilities Founders play a key role in the very early stages of a startup and also in setting the culture for the organisation, but as the organisation grows more importance is given to the influence of employees. Measures like the number of current employees are broad representations of the startup's human capital and are correlated with subsequent startup investment [6, 3, 12]. Detailed analyses of staff human capital are not present in the literature but may be possible using data collected from sources like AngelList and LinkedIn.

A.1.2 Social Capital

Entrepeneurship revolves around opportunity discovery and realisation [35]. Opportunity discovery is only possible through the medium of social networks, so social capital is important. Social networks exist in many forms and contribute in different ways to social capital. These networks can be categorised in terms of the strength of their relationships: weak ties (e.g. social media) and strong ties (e.g. strategic alliances).

Social Influence Startups use social media to communicate with other parties including their customers, potential customers, the media, potential employees, and potential investors. Social media activity can be proxy for a startup's social influence. Startups use different social media platforms for different purposes. Presence and engagement (e.g. number of followers, number of likes, number of posts) on Facebook and Twitter are predictive of startup investment success [11, 6]. These platforms are likely to capture customer or potential customer interactions, which is an indicator of

market adoption. In addition, the number of followers on AngelList predicts startup investment success [3], probably because it captures potential employees and investors' interest.

Strategic Alliances Strategic alliances with other companies or institutions have the potential to alter the opportunities that startups can access. Biotechnology startups that have links to industry partners are able to IPO more quickly and at higher market valuations [39]. Startups with more downstream (e.g. manufacturing), but not upstream (e.g. research and development), alliances obtain significantly more venture capital financing than startups with fewer such alliances [5].

A.1.3 Structural Capital

Structural capital is the supportive intangible assets, infrastructure, and systems that enable a startup to function. Intellectual property and their proxy, patents, are a key component of structural capital for newly-formed startups. Structural capital also includes processes and systems but these are less fully-formed in startups than more stable companies.

Patent Filings Many startups develop innovative technologies to help them capture a new market or better capture an existing market. Entrepreneurs protect their ideas through patent filings. Patents are an indicator of the technological capability of the startup. Patents and patent filings affect the survival and investment success of biotechnolgoy startups [5, 19]. However, there may not be as strong a relationship for non-biotechnology startups (e.g. software) [16, 1] This might be because factors like speed-to-market dominate the protective properties of patent filings in the quicker-moving high-technology sector.

A.2 Investment Confidence

A.2.1 Third Party Validation

By their nature, startups are optimsitic about the effectiveness of new technologies and business models. Founders are also highly invested in their startups and therefore it is reasonable for investors to doubt their claims. Third party validation from credible sources like other investors, the media, and the government, may be factored into investors' decision-making process [20, 18].

Investment Record Intuitively, a track record of demand for investment is likely to be a strong signal of future likelihood of future investment. Average funding per round, number of investors per round, number of previous financing rounds and total prior funding raised all predict future likelihood of investment [1, 6, 13, 19, 12].

Investor Reputation Funding from reputable investors sends a clear signal to potential investors that a startup is likely to be of high quality. Investors may believe they require less due diligence because it has been performed by another investor. Startups that receive their initial funding round from a prominent investor are more likely to survive and receive higher valuations in initial public offerings [18]. Followers on AngelList and previous co-investors predict the likelihood of an investor's portfolio startups raising additional rounds successfully [3, 41].

Media Coverage Media coverage provides legitimacy and credibility to startups. Media attention for startups affects the perceived valuation of well-informed experts like venture capitalists [28]. This also translates to increased investment success [6]. There are a few possible explanations for this. First, media coverage signals public interest which might positively influence other stakeholders like customers, employees, etc. Second, new information become widely available which reduces perceived information assymetry.

Awards and Grants Governments and other startup ecosystem supporters often run competitions, grant processes and awards to recognise and celebrate startups. Not only do awards and grants raise the profile of startups but they also indicate third-party validation. Interestingly, there is some evidence that government grants are positively associated with startup investment but awards may have a negative effect [1]. Perhaps this suggests that higher quality startups focus on more critical activities (e.g. raising funds, filing for patents).

A.2.2 Historical Performance

Startup performance is challenging to measure because there are no standardised reporting formats and the availability of data varies wildly. Capturing the multi-dimensionality of startup performance requires the use of multiple measures [42], however, most studies are only able to utilise simplistic performance metrics like survival time [30, 37, 17].

Financial Performance Despite being intuitive, there is little evidence of a relationship between startup financial performance and future investment success. This is because it tends to be difficult to access valid, accurate and complete financial performance measures (e.g. profit, revenue). This information is considered by startups as private and confidential and unlike public companies, private companies are not required to make financial disclosures. Proprietary databases can provide some data on private companies but commercial licenses are expensive and have poor coverage of early-stage companies [4].

Non-Financial Performance With a paucity of financial information available, researchers have looked for other measures of startup performance. Survival time is the most commonly studied startup performance metric despite the coarseness of the measure [37, 3, 16]. There are a few possible explanations for this. One explanation is that startups have such a high failure rate and long time to profitability that many won't ever report any other meaningful performance metrics [32].

A.2.3 Contextual Cues

Startups do not exist in isolation but are rather a product of their context. Investors must consider the performance of a startup's competitors, their local economy and the broader economy when evaluating the reasonableness of signals of startup potential.

Competitor Performance Startups are involved in almost every industry. However, startups across industries have very different requirements, trajectories and measure their performance in different ways. Comparing startups across industries does not necessarily provide a clear view as to whether the potential of a firm is remarkable, likely errant, or within normal ranges. Accordingly, industry classification has been found to be a key determinant of startup investment [34, 13, 16].

Local Economy Headquarters location is a key indicator of startup investment success [6, 13, 16]. A clear example of this effect is Silicon Valley, a location known for producing an outsized number of successful startups. Silicon Valley provides a focal point for engineering talent, previously successful entrepreneurs, and venture capital firms. Therefore, we might expect different signs of startup potential for Silicon Valley startups compared to those in locations where development and traction are more difficult to attain.

Broader Economy Although startups are less affected by broader economic trends than larger, well-established companies economic challenges have a knock-on effect for startup investment. The Global Financial Crisis led to a 20% decrease in the average amount of funds raised by startups per funding round, disproportionately affecting later-stage funding rounds. Therefore, when comparing startups of different ages, these sort of shocks have key implications for assessing what is a normal trajectory. This may explain why the year a startup is founded can influence startup investment [13, 19].

APPENDIX B

Data Sources

B.1 Databases

Databases play a critical role in understanding the startup ecosystem, aggregating information about startups, investors, media and trends. Most startup databases are closed systems that require commercial licenses (e.g. CB Insights, ThomsonOne, Mattermark). CrunchBase and AngelList are two crowd-sourced and free-to-use alternatives. AngelLists primary function is as an equity crowdfunding platform but it has a data-sharing agreement with CrunchBase which results in significant overlap between the two sources. CrunchBase and AngelList provide free Application Program Interfaces (API) for academic use. Crawlers can be developed to traverse these APIs and collect data systematically. The advantages of crawlers are that they can selectively collect data from nodes with specific attributes, collect random samples, or traverse the data source indefinitely, updating entries as new data becomes available. CrunchBase also provides pre-formatted database snapshots which allows easier access to the data set. The crowd-sourced nature of CrunchBase and AngelList has advantages and limitations . The key advantages are that access to the databases is free and the dataset is relatively comprehensive. The limitations are that both CrunchBase and AngelList have relatively sparse profiles (i.e. limited depth), particularly for unpopular startups. Both CrunchBase and AngelList also have error-checking provisions (including machine reviews and social authentication) to prevent and remediate inaccurate entries but there is still a greater chance for error. Comparing CrunchBase and AngelLIst, CrunchBase tends to have more comprehensive records of funding rounds [11] and media coverage but AngelList also has a social network element where users can ‘follow each other - in a similar way to Twitter.

B.2 Social Networks

Social networks provide an interesting perspective into the process of opportunity discovery and capitalisation that characterises entrepreneurship. Two social networks studied in detail in entrepreneurship research are LinkedIn and Twitter. LinkedIn is a massive professional social network often used in studies of entrepreneurship for measures of employment, education and weak social links. These measures are difficult to collect elsewhere. In addition, LinkedIn can provide a measure of the professional influence of founders and investors. Unfortunately, as of May 2015, the LinkedIn API no longer allows access to authenticated users' connection data or company data [40], making it difficult to use for social network analyses. Twitter is a massive social networking and micro-blogging service which is studied in entrepreneurship research because it is used by founders, investors, and customers to quickly communicate and broadcast. Twitter is a directed network where users can follow other users without gaining their permission to do so. Twitter's public API provides access to social network topological features (e.g. who follows who) and basic profile information (e.g. user-provided descriptions). However, Twitter's API only provides Tweets published within the last 7 days and access to historical Twitter data requires a commercial license [29].

B.3 Other Sources

While startup databases and social networks provide a variety of information on startups, there are two important areas that they do not cover: patent filings and financial performance. Startups often file patents to apply for a legal right to exclude others from using their inventions. In 2015, the US Patents Office (USPTO) launched PatentsView, a free public API to allow programmatic access to their database. PatentsView holds over 12 million patent filings from 1976 onwards [33]. The database provides comprehensive information on patents, their inventors, their organisations, and locations. It may be difficult to match identities across PatentsView to other data sources because registered company names (as in PatentsView) are not always the same as trading names (as elsewhere). Finding other information on startups, like financial information, is difficult. Unlike public companies, private companies are not required to file with the United States Securities and Exchange Commission (or international equivalent). Proprietary databases provide some data on private companies but commercial licenses are prohibitively expensive and have poor coverage of early-stage companies. PrivCo is one of few commercial data sources for private company business and financial

intelligence. PrivCo focuses its coverage on US private companies with at least $50-100 million in annual revenues but also has some coverage on smaller but high-value private companies (like startups) [4].

APPENDIX C

Classification Algorithms

C.1 Naive Bayes

Naive Bayes is a simple generative learning algorithm. It is a Bayesian Network that models features by generating a directed acyclic graph, with the strong (naive) assumption that all features are independent. While this assumption is generally not true, it simplifies estimation which makes Naive Bayes more computationally efficient than other learning algorithms. Naive Bayes can be a good choice for data sets with high dimensionality and sparsity as it estimates features independently. Naive Bayes sometimes outperforms more complex machine learning algorithms because it is reasonably robust to violations of feature independence [22]. However, Naive Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features [27]. Naive Bayes was used alongside Logistic Regression, Decision Trees and Support Vector Machines to predict success in equity crowdfunding campaigns on the AngelList data set [6]. None of these models performed well. The algorithm that best predicts startup investment was Naive Bayes with a Precision of .41 and Recall of .19, which means only 19% of funded startups were classified correctly by the model. The author suggests the poor performance of their algorithms is caused by features not captured in their data set relating to Intellectual Capital, Third Party Validation and Historical Performance. These features will be included in this study.

C.2 Logistic Regression

Regression is a class of statistical methods that investigates the relationship between a dependent variable and a set of independent variables. Logistic regression is regression where the dependent variable is discrete. Like linear regression, logistic regression optimises an equation that multiplies each input by a coeffi-

cient, sums them up, and adds a constant. However, before this optimisation takes place the dependent variable is transformed by the log of the odds ratio for each observation, creating a real continuous dependent variable on a logistic distribution. A strength of Logistic Regression is that it is trivial to adjust classification thresholds depending on the problem (e.g. in spam detection [15], where specificity is desirable). It is also simple to update a Logistic Regression model using online gradient descent, when additional training data needs to be quickly incorporated into the model (incremental learning). Logistic Regression tends to underperform against complex algorithms like Random Forest, Support Vector Machines and Artificial Neural Networks in higher dimensions [9]. This underperformance is observed when Logistic Regression is applied to startup investment prediction tasks [6, 7]. However, weaker predictive performance has not prevented Logistic Regression from being commonly used. Its simplicity and ease-of-use means it is often used without justification or evaluation [16].

C.3 K-Nearest Neighbours

K-Nearest Neighbours is a common lazy learning algorithm. Lazy learning algorithms do not produce explicit general models, but compare new instances with instances from training stored in memory. K-Nearest Neighbours is based on the principle that the instances within a data set will exist near other instances that have similar characteristics. K-Nearest Neighbours models depend on how the user defines distance between samples; Euclidean distance is a commonly used metric. K-Nearest Neighbour models are stable compared to other learning algorithms and suited to online learning because they can add a new instance or remove an old instance without re-calculating [22]. A shortcoming of K-Nearest Neighbour models is that they can be sensitive to the local structure of the data and they also have large in-memory storage requirements. K-Nearest Neighbours was compared to Artificial Neural Networks to predict firm bankruptcy [2]. K-Nearest Neighbours is attractive in bankruptcy prediction because it can be updated in real-time. By optimising feature weighting and instance selection, the authors improved the K-Nearest Neighbours algorithm to the extent that it outperformed the Artificial Neural Networks.

C.4 Decision Trees

Decision Trees use recursive partitioning algorithms to classify instances. Each node in a Decision Tree represents a feature in an instance to be classified, and

each branch represents a value that the node can assume. Methods for finding the features that best divide the training data include Information Gain and Gini Index [22]. Decision Trees are close to an "off-the-shelf" learning algorithm. They require little pre-processing and tuning, are interpretable to laypeople, are quick, handle feature interactions and are non-parametric. However, Decision Trees are prone to overfitting and have poor predictive power [10]. These shortcomings are addressed with pruning mechanisms and ensemble methods like Random Forests, respectively. Decision Trees were compared with Naive Bayes and Support Vector Machines to predict investor-startup funding pairs using CrunchBase social network data [24]. Decision Trees had the highest accuracy and are desirable because their reasoning is easily communicated to startups.

C.5 Random Forests

Random Forests are an ensemble learning technique that constructs multiple Decision Trees from bootstrapped samples of the training data, using random feature selection [8]. Prediction is made by aggregating the predictions of the ensemble. The rationale is that while each Decision Tree in a Random Forest may be biased, when aggregated they produce a model robust against over-fitting. Random Forests exhibit a performance improvement over a single Decision Tree classifier and are among the most accurate learning algorithms [10]. However, Random Forests are more complex than Decision Trees, taking longer to create predictions and producing less interpretable output. Random Forests were used to predict private company exits using quantitative data from ThomsonOne [7]. Random Forests outperformed Logistic Regression, Support Vector Machines and Artificial Neural Networks. This may be because the data set was highly sparse, and Random Forests are known to perform well on sparse data sets [8].

C.6 Support Vector Machines

Support Vector Machines are a family of classifiers that seek to produce a hyperplane that gives the largest minimum distance (margin) between classes. The key to the effectiveness of Support Vector Machines are kernel functions. Kernel functions transform the training data to a high-dimensional space to improve its resemblance to a linearly separable set of data. Support Vector Machines are attractive for many reasons. They have high predictive power [10], theoretical limitations on overfitting, and with an appropriate kernel they work well even when data is not linearly separable in the base feature space. Support Vector

Machines are computationally intensive and complicated to tune effectively (compared to Random Forests, for example). Support Vector Machines were compared with back propagated Artificial Neural Networks in predicting the bankruptcy of firms using data provided by Korea Credit Guarantee Fund [36]. Support Vector Machines outperformed Artificial Neural Networks, possibly because of the small data set.

C.7 Artificial Neural Networks

Artificial Neural Networks are a computational approach based on a network of neural units (neurons) that loosely models the way the brain solves problems. An Artificial Neural Network is broadly defined by three parameters: the interconnection pattern between the different layers of neurons, the learning process for updating the weights of the interconnections, and the activation function that converts a neuron's weighted input to its output activation. A supervised learning process typically involves gradient descent with back-propagation [31]. Gradient descent is an optimisation algorithm that updates the weights of the interconnections between the neurons with respect to the derivative of the cost function (the weighted difference between the desired output and the current output). Back-propagation is the technique used to determine what the gradient of the cost function is for the given weights, using the chain rule. Artificial Neural networks tend to be highly accurate but are slow to train and require significantly more training data than other machine learning algorithms. Artificial Neural Networks are also a black box model so it is difficult to reason about their output in a way that can be effectively communicated. Artificial Neural Networks are rarely applied to startup investment or performance prediction because research in this area typically uses small and low-dimensional data sets. As one author puts it "More complex classification algorithmsartificial neural networks, Restricted Bolzmann machines, for instancecould be tried on the data set, but marginal improvements would likely result." [6]. However, this study will address these issues so Artificial Neural Networks may be more competitive.

Bibliography

[1] Ahlers, G. K., Cumming, D., Gunther, C., and Schweizer, D. "Signaling in equity crowdfunding". In: *Entrepreneurship Theory and Practice* 39.4 (2015), pp. 955–980.

[2] Ahn, H. and Kim, K.-j. "Using genetic algorithms to optimize nearest neighbors for data mining". In: *Annals of Operations Research* 163.1 (2008), pp. 5–18.

[3] An, J., Jung, W., and Kim, H.-W. "A Green Flag over Mobile Industry Start-Ups: Human Capital and Past Investors as Investment Signals". In: *PACIS 2015 Proceedings*. AIS Electronic Library, 2015.

[4] Artemchik, T. "PrivCo". In: *Journal of Business & Finance Librarianship* 20.3 (2015), pp. 224–229.

[5] Baum, J. A. and Silverman, B. S. "Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups". In: *Journal of Business Venturing* 19.3 (2004), pp. 411–436.

[6] Beckwith, J. "Predicting Success in Equity Crowdfunding". Unpublished thesis. Joseph Wharton Research Scholars. Available at `http://repository.upenn.edu/joseph_wharton_scholars/25`. 2016.

[7] Bhat, H. and Zaelit, D. "Predicting private company exits using qualitative data". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Huang, J., Cao, L., and Srivastava, J. Vol. 6634. Lecture Notes in Computer Science. Berlin: Springer, 2011, pp. 399–410.

[8] Breiman, L. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[9] Caruana, R., Karampatziakis, N., and Yessenalina, A. "An empirical evaluation of supervised learning in high dimensions". In: *Proceedings of the 25th International Conference on Machine learning*. ACM. 2008, pp. 96–103.

[10] Caruana, R. and Niculescu-Mizil, A. "An empirical comparison of supervised learning algorithms". In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM. 2006, pp. 161–168.

[11] Cheng, M., Sriramulu, A., Muralidhar, S., Loo, B. T., Huang, L., and Loh, P.-L. "Collection, exploration and analysis of crowdfunding social networks". In: *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web*. ACM. 2016, pp. 25–30.

[12] Conti, A., Thursby, M., and Rothaermel, F. T. "Show Me the Right Stuff: Signals for High-Tech Startups". In: *Journal of Economics & Management Strategy* 22.2 (2013), pp. 341–364.

[13] Croce, A., Guerini, M., and Ughetto, E. "Angel Financing and the Performance of High-Tech Start-Ups". In: *Journal of Small Business Management* (2016).

[14] Fried, J. M. and Ganor, M. "Agency costs of venture capitalist control in startups". In: *New York University Law Review* 81 (2006), p. 967.

[15] Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*. Vol. 1. Berlin: Springer, 2001.

[16] Gimmon, E. and Levie, J. "Founder's human capital, external investment, and the survival of new high-technology ventures". In: *Research Policy* 39.9 (2010), pp. 1214–1226.

[17] Gloor, P. A., Dorsaz, P., Fuehres, H., and Vogel, M. "Choosing the right friends–predicting success of startup entrepreneurs and innovators through their online social network structure". In: *International Journal of Organisational Design and Engineering* 3.1 (2013), pp. 67–85.

[18] Hochberg, Y. V., Ljungqvist, A., and Lu, Y. "Whom you know matters: Venture capital networks and investment performance". In: *The Journal of Finance* 62.1 (2007), pp. 251–301.

[19] Hoenen, S., Kolympiris, C., Schoenmakers, W., and Kalaitzandonakes, N. "The diminishing signaling value of patents between early rounds of venture capital financing". In: *Research Policy* 43.6 (2014), pp. 956–989.

[20] Hsu, D. H. and Ziedonis, R. H. "Patents As Quality Signals For Entrepreneurial Ventures." In: *Academy of Management Proceedings*. Vol. 2008. 1. Academy of Management. 2008, pp. 1–6.

[21] Ingley, C. B. and McCaffrey, K. "Effective governance for start-up companies: regarding the board as a strategic resource". In: *International Journal of Business Governance and Ethics* 3.3 (2007), pp. 308–329.

[22] Kotsiantis, S. "Supervised Machine Learning: A Review of Classification Techniques". In: *Informatica* 31.3 (2007).

[23] Kuhn, M. and Johnson, K. *Applied predictive modeling*. Springer, 2013.

[24] Liang, Y. E. and Yuan, S.-T. D. "Predicting investor funding behavior using crunchbase social network features". In: *Internet Research* 26.1 (2016), pp. 74–100.

[25] Mitchell, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

[26] National Venture Capital Association. *2016 National Venture Capital Association Yearbook*. `http://www.nvca.org/?ddownload=2963`. Online; accessed 06 Nov 2016. Mar. 2016.

[27] Niculescu-Mizil, A. and Caruana, R. "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 625–632.

[28] Petkova, A. P., Rindova, V. P., and Gupta, A. K. "No news is bad news: Sensegiving activities, media attention, and venture capital funding of new technology organizations". In: *Organization Science* 24.3 (2013), pp. 865–888.

[29] Puschmann, C. and Burgess, J. "The politics of Twitter data". In: (2013).

[30] Raz, O. and Gloor, P. A. "Size really matters-new insights for start-ups' survival". In: *Management Science* 53.2 (2007), pp. 169–177.

[31] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. "Learning representations by back-propagating errors". In: *Cognitive Modeling* 5.3 (1988), p. 1.

[32] Sahlman, W. *Risk and reward in venture capital*. 2010.

[33] Schultz, L. A. "Preliminary Patent Searches: New and Improved Tools for Mining the Sea of Information". In: *Colo. Law.* 45 (2016), p. 55.

[34] Shan, Z., Cao, H., and Lin, Q. "Capital Crunch: Predicting Investments in Tech Companies". Unpublished thesis. Stanford University. Available at `http://www.zifeishan.org/files/capital-crunch.pdf`. 2014.

[35] Shane, S. and Venkataraman, S. "The promise of entrepreneurship as a field of research". In: *Academy of Management Review* 25.1 (2000), pp. 217–226.

[36] Shin, K.-S., Lee, T. S., and Kim, H.-j. "An application of support vector machines in bankruptcy prediction model". In: *Expert Systems with Applications* 28.1 (2005), pp. 127–135.

[37] Song, Y. and Vinig, T. "Entrepreneur online social networks–structure, diversity and impact on start-up survival". In: *International Journal of Organisational Design and Engineering* 2.2 (2012), pp. 189–203.

[38] Strobl, C., Malley, J., and Tutz, G. "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." In: *Psychological Methods* 14.4 (2009), p. 323.

[39] Stuart, T. E., Hoang, H., and Hybels, R. C. "Interorganizational endorsements and the performance of entrepreneurial ventures". In: *Administrative Science Quarterly* 44.2 (1999), pp. 315–349.

[40] Trachtenberg, A. *Changes to our Developer Program*. Ed. by Linkedin.com. `https://developer.linkedin.com/blog/posts/2015/developer-program-changes`. Online; accessed 18 05 2015. Feb. 2015.

[41] Werth, J. C. and Boeert, P. "Co-investment networks of business angels and the performance of their start-up investments". In: *International Journal of Entrepreneurial Venturing* 5.3 (2013), pp. 240–256.

[42] Wiklund, J. and Shepherd, D. "Entrepreneurial orientation and small business performance: a configurational approach". In: *Journal of Business Venturing* 20.1 (2005), pp. 71–91.

[43] Yu, Y. and Perotti, V. "Startup Tribes: Social Network Ties that Support Success in New Firms". In: *Proceedings of 21st Americas Conference on Information Systems*. 2015.

[44] Yuan, H., Lau, R. Y., and Xu, W. "The determinants of crowdfunding success: A semantic text analytics approach". In: *Decision Support Systems* 91 (2016), pp. 67–76.