Learning the factors that influence startup investment success

W.M.R. Shelton

This report is submitted as partial fulfilment
of the requirements for the Honours Programme of the
School of Computer Science and Software Engineering,
The University of Western Australia,
2016

Abstract

This is the abstract.

Keywords: keyword, keyword
CR Categories: category, category

Acknowledgements

These are the acknowledgements.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

CHAPTER 2

Literature Review

Technological advances have made launching a startup more accessible than ever before [48]. Billions of consumers can be accessed through the Internet and launching a startup can be done from a bedroom. However, startups remain competitive and risky endeavours. Technology startups can be unprofitable for years so entrepreneurs look for incubators, accelerators, angel investors and venture capitalists to support them through this developmental period. Aside from funding, investors hold other key resources (e.g. information, networks) that accelerate startup growth [20]. Investors act as scouts, able to identify the potential of new startups, and as coaches, able to help startups realise that potential [7].

It is important for startups to continue to win over investors throughout their development, but this is not a trivial task. Raising funding rounds from investors can be challenging and time-consuming, as investors find it difficult to quickly evaluate startups as investment opportunities. Investors spend time seeking and evaluating signals of a startup's underlying quality because clear metrics of performance often do not exist or are difficult to capture [43]. The growing popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of a desire for a more efficient assessment of startup potential. By the end of 2014, over 1,200 investment organisations (including 624 venture capital firms) were members of CrunchBase's Venture Program, mining CrunchBase's startup data to help inform their investment decisions [38].

Venture capital investment comes with trade-offs for startups. About half of venture capital-backed startups end in complete liquidation [27]. Investors are protected from these losses because the minority of their investments that do survive have outsized returns: 85% of venture capital returns come from just 10% of investments [41]. As the venture capital industry matures, investors optimise for this extreme risk-reward trade-off by pushing startups to grow rapidly, frequently raise follow-on funding rounds and make quick, centralised decisions [23]. The rapid growth demanded by venture capital investors is generally incompatible with public company structures, due to strict reporting and compliance require-

ments [52]. Accordingly, venture-capital backed startups are delaying their Initial Public Offerings (IPO): Time to IPO has doubled in the past 20 years [13].

Startups remaining privately-held for longer has the effect of shifting value creation to the private markets. For example, Microsoft's market capitalisation grew 500-fold following its IPO in 1986 [34], but for Facebook to grow to the same extent since its IPO in 2012 its valuation would exceed the capitalisation of the global equity market [39]. Venture capital funding for late-stage privately-held startups is approaching all-time highs as investors enter the private markets [13]. It is important to understand how the factors that influence venture capital investment change throughout a startup's development. There is a clear gap in the academic literature in learning how the factors that influence startup investment change throughout a startup's development. Previous approaches to learning the factors that influence venture capital investment in startups have common weaknesses. This study will address these weaknesses in three ways:

Large Sample Size Prior work are largely restricted in sample size. Most studies in this area have sampled fewer than 500 startups [2, 17, 26, 21], or between 500 and 2,000 startups [28, 53, 5, 51, 19], and only a few have used large scale samples (more than 100,000 startups) [42, 16]. Abundant empirical evidence has suggested that the size of training data eventually becomes more critical than the sophistication of algorithms themselves or even careful feature selection [14]. Large open databases (e.g. CrunchBase, AngelList) and social networks (Twitter, LinkedIn) offer larger samples than those generally studied in previous works. We expect that using data collected from these sources will lead to the discovery of additional features and higher accuracy in startup investment prediction.

Developmental Focus Prior work has focused primarily on early-stage investment in startups, primarily equity crowdfunding [8, 2, 16, 54] and angel investing [19]. The functions and objectives of startups change through their development [33]. We expect that the signals that attract investment in these companies will similarly change over time.

Rich Features Prior work has focused on factual company details (e.g. the headquarters' location, the age of the company, the number of founders) for startup investment predictive models [8, 21, 26]. Semantic text features (e.g. patents and media) [28, 54, 50] and social network features (e.g. co-investment networks) [49, 51, 16, 53] may also predict startup investment. We expect that

developing a comprehensive model that includes semantic text and social network features alongside factual company features in could lead to better startup investment prediction.

This study will develop software that collects and processes information on startups to predict their likelihood of raising investment at different stages in their development. If successful, this study has the potential for scholarly, policy and firm-specific implications. We propose a conceptual framework for startup investment, based on work by Ahlers and colleagues (2015) [2] and Baum and Silverman (2004) [7]. Our conceptual framework models startup investment success as a product of two factors: venture quality and investment confidence. We will test this framework with respect to startup development, using cross-sectional and longitudinal analyses. We hope that this study provides interesting insights for entrepreneurs, policy makers, and investors and improve their understanding of the determinants of startup investment, especially for later-stage startups. Ultimately, we hope that this study encourages greater investment in startups.

The paper proceeds as follows. The next section explores theoretical models of technology startups and startup investment (Section 1). Thereafter, we review empirical evidence of features linked to startup investment (Section 2). We then determine how to collect the data to test those features (Section 3) and evaluate machine learning algorithms to find those that suit this startup investment prediction task (Section 4). The final section summarises our main results and concludes.

2.1 Theoretical Background

Startups are an important means for the commercialisation of new technological discoveries. Often, startups introduce disruptive technologies and perform the role of Schumpeterian entrepreneurship, or "creative destruction", in the economy [46]. the factors that drive their performance are important to understand. For startups, rapid growth is generally an indication of wide market acceptance of their products or services. However, growth is difficult to achieve, and most startups remain small or become bankrupt [27]. In this section, we discuss the theory behind the drivers of entrepreneurial growth and explore the role of external investment. We propose a conceptual framework for startup investment success that will be evaluated in the current study.

2.1.1 Technology Startups

Steve Blank, entrepreneur-mentor and author, defines a startup as "an organization formed to search for a repeatable and scalable business model." [10]. In this case "search" is intended to differentiate established late-stage startups from traditional small businesses, such as restaurants. Often, startups are based on applying technologies to problems in ways that are significantly distinct from how the problems are currently solved. In this section, we discuss the entrepreneurial theory that underpins startups including the startup development lifecycle, and the factors that contribute to their growth and performance.

2.1.1.1 Developmental Lifecycle

Entrepreneurship is a process and the functions and objectives of startups change dramatically through their development [33, **?**]. The startup development lifecycle can be generally divided into multiple stages, though there is substantial variance between startups. First, there exists a period in which the startup founders identify a need or problem that they wish to solve. Second, the founders develop an idea and decide to create a company. These stages may be in rapid succession or drawn out, depending on the intensiveness of the solution development process. In these stages, where the firm has not been created yet, the knowledge, experience and attitudes embedded in the founders are the most critical features. The third stage is the gestation period. Once the company is created, the entrepreneurs must make a considerable effort to adjust the company to market conditions, and most importantly, to build a successful organization. Generally, it is at this stage that startup founders will raise one or more rounds of external investment to help accelerate their growth. The outcome of this turbulent gestation process is a consolidation period of some sort. The consolidation process of a start-up firm can take the form of a business failure, reversion (to another problem or solution), decline, stability, and growth.

2.1.1.2 Determinants of Performance

A review of the literature has found that the determinants of startup performance are often categorised into three fundamental elements: human capital, social capital and structural capital [7, 2, **?**].

Human Capital Newly-formed startups have few financial resources and are still searching for a viable business model and customer demand, so human capital

is a dominant resource. Human capital has been defined as "a stock of personal skills that economic agents have at their disposal" [?]. In an entrepreneurial context, human capital is related to identifying and exploiting business opportunities [?], defining and realizing a venture's strategy, acquiring additional resources, and building a positive basis for future learning. A meta-review of the effect of founders' human capital showed that industry-related experience, education and founders' team compatibility are significant factors contributing to new venture survival [26]. Accordingly, venture capitalists indicate that experience and management skills are among their most important selection criteria [?].

Social Capital Social capital is the value derived from the utilisation of social networks [?]. Startups are often dependent upon external resources in the early stages of their development because they tend to take time to become profitable. Entrepreneurs require valuable resources such as information, advice, finance, skills and labour when launching startups to be able to realise entrepreneurial opportunities [?] and it is social networks that provide the media for those resources to be sourced and transmitted. Social capital has been studied as a potential determinant of various entrepreneurial performance metrics, including survival time [?, 45], venture capital raised [?] and revenue generated [?, 22]. Entrepreneurs that are centrally embedded in their social networks are more likely to access important resources [?, 45]. Being a key influencer or aligning oneself with key influencers can increase the quality of an entrepreneur's connection [?]. Such works suggest that strengthening and maintaining social networks plays an essential role in the performance of startups.

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. Innovation is a key determinant of firm survival. It can simultaneously allow new firms to enter the market while helping established firms secure their competitive positions and thus their survival A patent constitutes a legal right to exclude others from using an invention. As such, patents support the appropriation of returns from innovative activities and facilitate cooperation and bargaining with business partners. Indeed, patent ownership has been shown to be correlated with company valuation [?]. Furthermore, patent ownership correlates with the business performance of start-ups in terms of asset growth [?], short time to initial public offering (IPO) [?], and an increased likelihood of survival after IPO [?]. Accordingly, there is also strong evidence for a positive relationship between patent filings and startup investment success [?].

2.1.2 Startup Investment

Startups often seek to obtain external risk investment from angel investors, venture capital and private equity firms. Given the uncertainty and imperfect information that typically surround these investment opportunities, the startup investment decision-making process is difficult for startups and investors alike. In this section, we discuss what motivates startups to seek external risk investment, how investors discover and make investment decisions, and outline the factors that drive investor confidence.

2.1.2.1 Benefits of Investment

Many obstacles confront young companies. Startups have little operating experience, and frequently operate using immature and unrefined routines, with little knowledge of their environment, and poor working relationships with customers and suppliers. In addition to this inexperience, startups often require substantial resources to fund speculative development projects, while remaining unprofitable sometime into the future. This uncertainty is compounded for firms established to pursue commercial applications of new technologies [**?**]. In this respect, external risk investment is considered by both academics and practitioners as one of the key drivers of the success of entrepreneurial firms and it is typically highly desirable for early-stage startups. Venture capital-backed firms have been shown to grow faster, patent more, have higher productivity and are more likely to go public [**?**]. This effect may be explained in three ways: first, that venture capitalists are effective scouts of potentially successful startups; second, that venture capitalists provide resources that help startups become more successful; and third, that the act of venture capitalists investing in a startup is a signal that encourages other third parties (e.g. future employees, customers, investors) to be more confident in the potential of the startup.

2.1.2.2 Funding Rounds

TODO

2.1.2.3 Investment Approaches

Investors use software to assist them in discovering, evaluating and predicting the performance of startups. It is likely that most this software is not disclosed. In 2008, the well-funded startup YouNoodle announced that they had developed software that could predict the future valuation of startups based on analysis

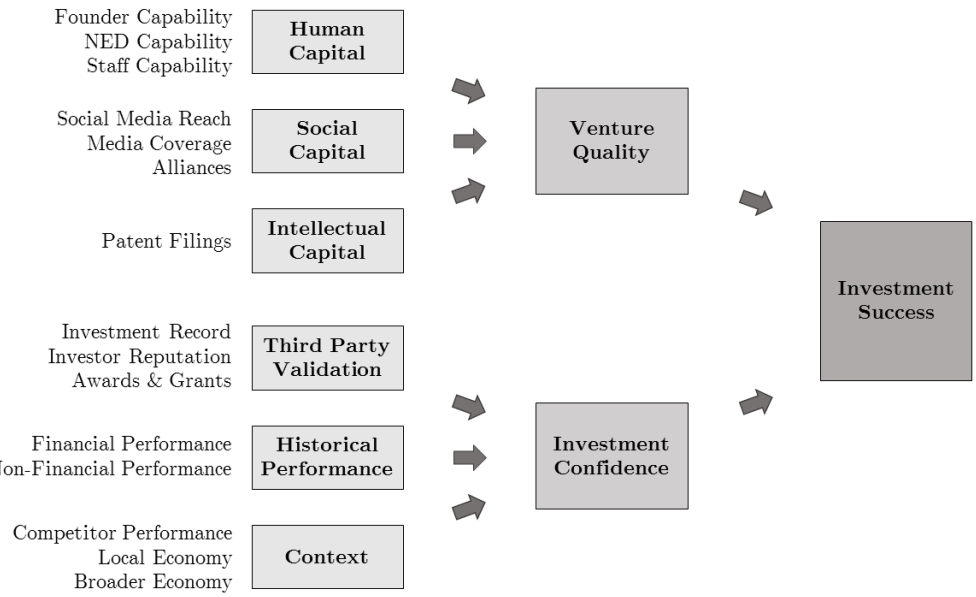
of their founding teams [6]. In 2010, the venture capital firm Kleiner Perkins Caulfield Byers announced that they had developed software called Dragnet that digests App Store data, AngelList entries and Twitter mentions (amongst other data), to surface early-stage startups [25]. Investors use some combination of two approaches to evaluate startup potential: extrapolation of current performance metrics and prediction based on underlying determinants of performance. Dragnet directly evaluates current metrics of startup performance (e.g. app downloads, viral momentum etc.) [25] while YouNoodle analyses determinants of startup performance (in this case, the human capital of the founding team) [6]. Both approaches have strengths and weaknesses. Dragnet's method of evaluating current performance metrics is easier to implement but YouNoodle's method of evaluating determinants has the potential to be more powerful and explanatory.

2.1.2.4 Investment Confidence

A key challenge of the startup investment process is informational asymmetry. Founders possess far more information than investors regarding a venture's prospects. Expecting founders to fully inform a potential investor is unrealistic, especially given that institutional knowledge is embedded in the skills and capabilities of the founders and may be unrecognized by the founders themselves [**?**]. Potential investors try to evaluate the unobservable characteristics of venture quality by interpreting the signals sent by entrepreneurs as well as potentially a company's attributes [**?**]. To get an understanding of the quality of a startup's signals, investors may look to other factors to corroborate the evidence like existing third party validation (e.g. previous investments), historical performance (e.g. profitability), and contextual cues (e.g. performance of competitors).

2.1.3 Proposed Framework

External risk investment is highly desirable and a critical successful factor for technology startups. However, our understanding of the factors that drive startup investment success is incomplete. We have discussed two key theoretical drivers for investors making startup investment decisions: first, understanding determinants of venture quality and second, qualifying the level of confidence they have in signals about that venture quality. Ahlers and colleagues (2015) [2] developed a conceptual framework for funding success on equity crowdfunding platforms. Their framework had two key factors: venture quality and level of uncertainty. The first factor is based on Baum and Silverman's (2004) [7] structure that suggests the key determinants of startup venture quality are human capital, social (alliance) capital, and intellectual (structural) capital. The second factor is based



on investors' confidence in their estimation of venture quality based on concerns about information asymmetries between themselves and founders.

We seek to generalise Ahlers and colleagues' [2] framework to other stages of the startup development lifecycle (beyond equity crowdfunding). While the first factor of Ahlers and colleagues' framework (venture quality) is generalisable to startups of all stages, Ahlers and colleagues operationalised the second factor with respect to whether startups offered an equity share in return for their crowdfunding, and whether they provided financial projections. These features are less generalisable to later-stage startups. We propose an extension of Ahlers and colleagues' framework that provides a more developed second factor. We describe investment confidence as a product of third party validation, historical performance and contextual information. Our proposed framework is depicted in Figure 1.

2.2 Feature Selection

In the previous section, we developed a conceptual framework relating venture quality and investor confidence to startup investment success. We seek to operationalise this conceptual framework into features that can be incorporated into our machine learning model. Table 1 shows a review of empirical evidence of features used in previous studies that have explored the startup investment process.

In this section, we describe each of these features, and outline conceptual and empirical evidence that justify their inclusion in our conceptual framework.

2.2.1 Feature Evaluation

We reviewed potential features for their inclusion under our proposed conceptual framework for use in our machine learning model. Our proposed conceptual framework is based on an earlier framework by Ahlers and colleagues (2015) [2]. Our proposed framework and feature set builds upon this work in multiple 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 and colleagues used one or two features to represent each factor in their model (e.g. "% Nonexecutive board" represents "Social (alliance) capital"), we performed a review of many different features used in this area and have performed a higher level 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 broken down into a number of features, data sources permitting.

2.3 Data Sources

Predicting startup investment is a complex 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. In Table 2 we have outlined the general characteristics of a selection of relevant data sources and how they can contribute to the features we're studying. In this section, we describe the desired 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.

2.3.1 Source Characteristics

Online data sources typically capture a wide spectrum of businesses and entrepreneurs with varying data quality. It is important to filter entities to some extent before further analyses are performed. For this study, we will restrict our

scope to companies that are based in the United States. The following sections describe common data sources used in entrepreneurship research, particularly data sources that focus on technology startups in the United States.

2.3.1.1 Surveys and Interviews

Researchers in the field of entrepreneurship have historically relied heavily on surveys and interviews for data collection [**?**]. Entrepreneurs may be more likely to provide private and confidential information in surveys and interviews than in other contexts. Information about human capital (e.g. founder, directors' and staff capabilities), strategic alliances, and financial performance may be difficult to capture elsewhere. The trade-off for access to these features is that surveys and interviews are generally time-consuming and costly to implement. While online surveys address some of these issues there is still an issue of motivating potential participants to contribute. Startup databases and social networks can be more efficient for data collection than surveys and interviews. For participants interacting with these sources, data collection is a secondary function so the participants do not require prompting from the researcher. Generally, the researcher can also collect the results from these sources automatically and at scale.

2.3.1.2 Startup Databases

Startup databases collect and store information about startups, investors, media coverage and trends. Most startup databases are closed systems that require expensive commercial licenses to use (e.g. CB Insights, ThomsonOne, Mattermark). CrunchBase and AngelList are two large crowd-sourced and free-to-use alternatives. CrunchBase and AngelList provide free Application Program Interfaces (API) for academic use. Dataset crawlers can be developed to traverse these APIs and collect data systematically. The advantages of such a crawler is that it can selectively collect data from nodes with specific attributes, or collect a random sample, or traverse the data source indefinitely, updating entries as new data becomes available. CrunchBase also provides public pre-formatted database snapshots which allows easier access to the dataset.

CrunchBase CrunchBase is an open online database of information about startups, investors, media coverage and trends, focusing on high-tech industry in the United States. It relies on its online community to edit most pages. CrunchBase is a comprehensive database, with almost complete coverage of startups and investors in the Internet sector, including the relationships between them

[4]. However, it has been noted that the CrunchBase corpus is sparse with many missing attributes [55]. CrunchBase has three provisions to prevent and remediate inaccurate crowd-sourced entries [1]. First, all users are required to authenticate their CrunchBase accounts with a social media account which allows CrunchBase to verify a user's identity. Second, every change goes through a machine review, which flags significant or questionable updates for moderation. Third, well-known startups have their editing privileges locked and require email verification.

AngelList AngelList is a promising new source of startup data, combining the functionality of an equity crowdfunding platform, a social networking site and an online startup database. As an equity crowdfunding platform, users create profiles for their startups on AngelList, and use the platform to attract investment. Investors use the platform to identify investment opportunities and can invest directly through AngelList, often alongside other investors in investment syndicates. AngelList is also an online startup database. It has a data-sharing agreement with CrunchBase which results in significant overlap between the two sources, though CrunchBase tends to have more comprehensive records of funding rounds [16]. Importantly, AngelList also tracks "startup roles" (e.g. founders, investors, employees) with a creation time, start time and end time. This means that, unlike CrunchBase, AngelList's networks can be re-created through time. For example, Britz et al. [12] analysed the formation of relationships on AngelList (looking at edge-creation time-stamps) to perform a longitudinal study of community growth and development.

2.3.1.3 Social Media

Social media platforms allow people to network and communicate online. As a side-effect, they also capture useful information about peoples' identities and relationships. There are many diverse social networking sites: Facebook, friendships; LinkedIn, professional relationships; Twitter, micro-broadcasts; Snapchat, ephemeral video. Two social media platforms have been focused in on for academic research on entrepreneurship: LinkedIn and Twitter.

LinkedIn LinkedIn is commonly used in online social network studies of entrepreneurship because it is a social network primarily used for professional networking [22, 45]. LinkedIn captures information about founders, directors and staff capability including past employment and education which is difficult to collect elsewhere. In addition, it can capture the professional social reach of

founders and investors. Unfortunately, as of May 2015, the LinkedIn API no longer allows access to authenticated users' connection data or company data [47], making it virtually impossible to use this site for social network analysis, without resorting to semi-automatic HTML parsing techniques (which is against the Terms of Service).

Twitter Twitter is a social networking site and micro-blogging site that is often used by entrepreneurs to promote their personal and business brands and rapidly share news and opportunities. Users can send and read public messages (called tweets) of 140-character length. 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 description). However, Twitter's public API only provides Tweets published within the last 7 days. To access historical Twitter data requires a commercial license to Gnip (also owned by Twitter).

2.3.1.4 Other Sources

Patent Filings Startups 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 [**?**]. The database provides comprehensive information on patents, their inventors, their organizations, and locations. One limitation might be difficulty matching identities across multiple data sources because registered company names (as reported in PatentsView) are not always the same as trading names (as reported in other data sources).

Financial Reports Finding private company financial information is difficult. Unlike public companies, private companies are typically not required to file with the United States Securities and Exchange Commission (or international equivalent). There are a few proprietary databases that provide this data but typically commercial licenses are prohibitively expensive. These databases also don't tend to cover early-stage companies. For example, PrivCo is a source for private company business and financial intelligence that covers over 500,000 private companies. PrivCo focuses its coverage on U.S. Major 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) [**?**].

2.3.2 Source Evaluation

We evaluated relevant data sources for their suitability to the current task, startup investment prediction. The startup databases CrunchBase and AngelList provide the most comprehensive set of features for this task. There are small differences between the features recorded in each. CrunchBase has slightly more coverage and tracks media mentions better but lacks AngelList's social network and timestamping. At least one startup database should be used in this study, but it is less significant which one is used. Of the other data sources reviewed, patent filings and Twitter are the most promising. The US Patent Office has a free public API with comprehensive patent information, though it could be difficult matching identities across multiple sources because registered company names are not always the same as trading names. Twitter provides social network topology and basic profile information through its free API but doesn't provide access to historical tweets. Other data sources are less promising because of access issues. Although Surveys and Interviews provide insights into human capital and past performance, manual data collection doesn't suit the scope of this study. LinkedIn cannot be easily collected now that the API is deprecated. Financial reports are too expensive to access for the purposes of this study.

2.4 Learning Algorithms

Machine learning is characterised by algorithms that can 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." [36]. Machine learning algorithms can be classified based on the nature of the feedback available to them: supervised learning, where the computer is presented with example inputs and desired outputs; unsupervised learning, where no labels are provided and the computer must find structure in its input; and reinforcement learning, where a computer 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 (basically, unsupervised classification). In this section, we describe the characteristics of the startup investment prediction task, review common ma-

chine learning algorithms, and ultimately determine which algorithms are most likely to suit the characteristics of this task.

2.4.1 Task Characteristics

Machine learning tasks are diverse and can be approached in many ways. For the current study, we will manipulate and combine the data collected from our data sources into a labelled data set appropriate for the application of supervised machine learning algorithms. Primary labels will be whether a startup received funding at each funding round, though measures of startup performance may also be investigated (e.g. survival time, exit). The key objective of machine learning algorithm selection is to find algorithms that make assumptions that are 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 Table 5 we have outlined some of the general characteristics of supervised learning tasks and identified the characteristics that are relevant to the current startup investment prediction task.

2.4.1.1 Data Set Properties

Missing Values Data sets often have missing values, where no data is stored for a feature of an observation. Missing data can occur because of nonresponse 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 CrunchBase and social networks, are typically sparse with missing entries despite their scale.

Mixed feature types Data sets can contain data with distinct primitive natures: real-valued, interval, counts, rank, binary, ordinal, categorical and multi-categorical types. The simplest way to handle mixed data types is to convert into a unified type (e.g. real-valued, binary). However, this process partially destroys type-specific information. We expect mixed-feature types in our dataset as we will be handling data from databases, social networks and semantic text analysis.

Irrelevant features Despite best efforts to only include features that have some theoretical relevance, most machine learning tasks will include irrelevant features. Irrelevant features are those that have no underlying relationship with classification. However, depending on the way they are handled they may affect

classification or slow down the learning algorithm. We expect irrelevant features in our dataset because our proposed framework includes features that haven't been tested in the literature.

Imbalanced classes 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 dataset is not dramatically imbalanced overall, but when looking at funding status for different funding rounds it is significantly imbalanced.

Small training set Machine learning techniques generally work better with more data [14]. Problems of small-data are numerous, but mainly revolve around high variance: over-fitting is harder to avoid, noise is an issue and outliers become more significant. Some machine learning techniques are better at dealing with these problems with others. The primary datasets that we're looking at are reasonably large (more than 100,000 startups) so this is not a relevant characteristic.

High dimensionality In extremely high dimensional data sets, the number of features is larger than the number of observations. When dimensionality increases to the extent, the volume of the feature space increases so fast that the available data becomes sparse. This phenomenon is known as the curse of dimensionality or "Hughes effect" [30]. We do not expect our data set to be highly dimensional and where possible will transform the data into concise features.

2.4.1.2 Desired Algorithm Properties

Predictive Power Predictive power is the ability of a machine learning algorithm to correctly classify new observations. If a model has no predictive power, then fundamentally the model is not representing the underlying statistical process being studied. For this reason, predictive power is a dominant characteristic. However, predictive power is far from the only important characteristic in machine learning selection. If algorithms provide similar predictive power, then other selection criteria become more significant. Predictive power can be evaluated in many different ways. For the purposes of this startup investment task, with imbalanced class distribution, we will be looking at metrics like Area under the Receiver-Operator Curve and the F1 Score. In particlar, we will be focusing on the predictive power for the positive classes.

Interpretability 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 important that we understand the determinants of the model so we are seeking algorithms that are highly interpretable.

Incremental Learning Incremental learning is where learning occurs dynamically whenever new observations are made and the algorithm adjusts what has been learned per the new observations. The key driver behind the need for incremental learning is when the underlying source generating the data is changing. It is plausible that, as a system, the drivers behind startup investment are changing over time.

Ease of Tuning Machine learning algorithms have hyperparameters that must be tuned to ensure the model does not overfit its dataset. Some algorithms have many hyperparameters and tuning can be a computationally expensive process. We are not under significant time-pressure for this study and our dataset is not prohibitively large so this characteristic is not especially relevant.

Computational Speed The amount of time and computational resources necessary to train a model varies a great deal between algorithms. Training time is often closely tied to predictive power. In addition, some algorithms are more sensitive to the number of data points than others. When time is limited it can drive the choice of algorithm, especially when the data set is large. For this task, with a relatively moderate data set and without significant time-pressure, this characteristic is not especially relevant.

2.4.2 Algorithm Characteristics

Supervised machine learning are algorithms that can reason about observations to produce general hypotheses, and then 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). The meta-review in Table 6 compares common supervised learning algorithms across general characteristics of the two domains mentioned in the previous section: assumptions about the structure of the problem and constraints of the desired solution. The following sections describe

each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in relevant applications.

2.4.2.1 Naive Bayes

Naive Bayes is a simple generative learning algorithm. It is a form of 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 means Naive Bayes is more computationally efficient than other learning algorithms. Naive Bayes can be a good choice for datasets with high dimensionality and sparsity as it estimates each feature independently. Naive Bayes has been found to sometimes outperform more complex machine learning algorithms because it is reasonably robust to violations of feature independence, at least with regards to classification [37]. However, Naive Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features [40]. 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 [8]. None of these models performed well. The algorithm that best predicted funded startups was Naive Bayes with a Precision of .41 and Recall of .19, which means that only 19% of funded startups were classified correctly by the model. The author suggests the poor performance of their algorithms is caused by insufficient features captured in their training set, missing features relating to Intellectual Capital, 3rd Party Validation or Historical Performance. These features are included in the theoretical framework proposed by the current study.

2.4.2.2 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 coefficient, 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 [24], where it is important that specificity is high). 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. Logistic Regression tends to underperform

against more complex algorithms like Random Forest, Support Vector Machines and Artificial Neural Nets in higher dimensions [14]. This underperformance is observed when Logistic Regression is applied to startup investment prediction tasks [8, 9]. However, weaker predictive performance hasn't prevented Logistic Regression from being commonly used. Its simplicity and ease-of-use means it is used more casually, often being used without justification or comparative evaluation of its use [26, 29].

2.4.2.3 K-Nearest Neighbours

K-Nearest Neighbours is a common lazy learning algorithm. Lazy learning algorithms do not perform explicit generalisation, but compare new instances with instances from training stored in memory. K-Nearest Neighbours is based on the principle that the instances within a dataset will exist near other instances that have similar characteristics [18]. 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 [31]. 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 [3]. 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 managed to improve the K-Nearest Neighbours algorithm to the point where it outperformed the Artificial Neural Network.

2.4.2.4 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 [35]. 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 [15]. These shortcomings have been 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 [32]. Decision Trees had the highest classification accuracy and the authors suggest they are particularly useful in this application because their reasoning is easily communicated to startups.

2.4.2.5 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 [11]. 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 that is robust against over-fitting. Random Forests exhibit a performance improvement over a single Decision Tree classifier and are among the most accurate learning algorithms [15]. 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 [9]. 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 [11].

2.4.2.6 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 typically high accuracy [15], theoretical guarantees on limiting overfitting, and with an appropriate kernel they can work well even if data is not linearly separable in the base feature space (though this is an issue with a linear kernel). Support Vector Machines are computationally intensive and relatively 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 [44]. Support Vector Machines were found to outperform Artificial Neural Networks at this task, especially because it was on a small data set.

2.4.2.7 Artificial Neural Networks

Artificial Neural Networks are a computational approach based on a network of neural units (neurons) that loosely models the way that 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 [**?**]. 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 very 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 have been rarely applied to startup investment or startup performance prediction tasks, probably because research in this area has used relatively small and low dimensional data sets. As one author put it "More complex classification algorithmsartificial neural networks, Restricted Bolzmann machines, for instancecould be tried on the data set, but marginal improvements would likely result." [8]. However, the current study seeks to address both of those issues so Artificial Neural Networks may be more relevant.

2.4.3 Algorithm Evaluation

We evaluated common supervised learning algorithms for their suitability to the current task, startup investment prediction. In Table 7 we produce a ranking based on cross-referencing the task characteristics with the characteristics of the common algorithms. While this gives us some directionality of fit, we hesitate to rule in or out algorithms purely based on this ranking. Algorithm selection is complex and preliminary empirical testing will provide clarity as to which algorithms should be used. In addition, larger training sets and good feature design tends to outweigh algorithm selection [14]. With those concessions in mind, we continue to review our findings. Our findings suggest that we should expect Random Forests, Support Vector Machines and Artificial Neural Nets to produce the highest classification accuracies. An ensemble of these high-performing methods may also provide an accuracy improvement, though at the cost of computational

speed and interpretability. Random Forests could be expected to slightly 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 might be preferable for early, exploratory analysis of the dataset.

2.5 Conclusion

A literature review was conducted to determine how to learn the factors that influence investment success for startups. First, we explored theoretical models of technology startups and startup investment (Section 1). Thereafter, we reviewed empirical evidence of features linked to startup investment (Section 2). We then determined how to collect the data to test those features (Section 3) and evaluated machine learning algorithms to find those that suit this startup investment prediction task (Section 4).

Venture capital funding for late-stage privately-held startups is approaching all-time highs as investors enter the private markets [13]. It is important to understand how the factors that influence venture capital investment change throughout a startup's development. There is a substantial research gap around accurately predicting startup investment success. Existing approaches in the literature were assessed to have three common limitations: small sample size [2, 17, 26, 21, 28, 53, 5, 51, 19], a focus on early-stage investment [8, 2, 16, 54, 19], and sparse use of features [2, 5, 16, 19, 51, 26]. Although individual studies addressed some of these limitations, none attempted to synthesise their findings into a standalone study and piece of software.

This study will develop software that collects and processes information on startups to predict their likelihood of raising investment at different stages in their development. If successful, this study has the potential for scholarly, policy and firm-specific implications. We propose a theoretical framework for startup investment, based on work by Baum& Silverman (2004) [7] and Ahlers and colleagues (2015) [2]. Our theoretical framework models startup investment success as a product of two factors: venture quality and investment confidence. We will test this framework with respect to startup development, using cross-sectional and longitudinal analyses. We hope that this study provides interesting insights for entrepreneurs, policy makers, and investors and improve their understanding of the determinants of startup investment, especially for later-stage startups. Ultimately, we hope that this study encourages greater investment in startups.

CHAPTER 3

Methods

CHAPTER 4

Results

CHAPTER 5

Discussion and Conclusion

APPENDIX A

Appendix

A.1 Feature Selection

A.1.1 Venture Quality

As startups change rapidly throughout their development, traditional assessments of company valuation (e.g. discounted cashflows, comparable valuation) are difficult to implement accurately. Instead, when valuing startup venture quality, we typically look at more fundamental drivers. These determinants of startup performance are often categorised into three elements: human capital, social capital and structural capital [7, 2, **?**].

A.1.1.1 Human Capital

Human capital is critical to early-stage startups that have few other 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 unique contribution to the human capital of the startup. Additionally, the human capital of each of these parties can generally be categorised three ways: their education, previous experience, and synergies as a team.

Founders' Capabilities Founders play multi-skilled 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, the education of founders is a key signal. The number of degrees attained by founders is predictive of success [8, 26] as has whether a founder has obtained an MBA [8]. In addition, past entrepreneurial experience seems to be a predictive factor [26] though there is some evidence to dispute this [42]. Finally, the number of founders seems to be correlated to startup success [8],

though the underlying relationship may be more nuanced, and could be related to the distribution of team skillset.

Non-Executive Directors' Capabilities TODO

Staff Capabilities TODO

A.1.1.2 Social Capital

Entrepeneurship revolves around opportunity discovery and realisation [**?**]. 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 have been found to be predictive of startup investment success [16, 8]. 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 has been found to predict startup investment success [5], probably because it captures potential employees and investors' interest.

Strategic Alliances TODO

A.1.1.3 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 other processes and systems but these have not been reviewed empirically in the literature, probably because they are more difficult to capture and operationalise.

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 been shown to affect the survival and investment success of biotechnolgoy startups [7, 28]. However, other studies have not shown as strong a relationship for non-biotechnology startups (e.g. software) [26, 2] This might be because factors like speed to market dominate the protective properties of patent filings in the quicker moving high-technology sector.

A.1.2 Investment Confidence

Investors face high information assymetry when assessing startup venture quality. Accordingly, they seek to get an understanding of the quality of the signals that they receive from startups. This may include reviewing third party validation (e.g. previous investments), historical performance (e.g. profitability), and contextual cues (e.g. performance of competitors).

A.1.2.1 Third Party Validation

TODO

Investment Record TODO

Investor Reputation Receiving funding from a highly reputable investor sends a clear signal to other potential investors that a startup is more likely to be of high quality. Investors may feel that they require less due diligence on the startup because another investor has already performed due diligence. Accordingly, studies show that startups that receive their initial funding round from a prominent investor are more likely to survive and receive higher valuations in initial public offerings [**?**]. A study showed that the number of followers an investor had on AngelList predicted the likelihood of their portfolio startups raising additional rounds successfully [5]. Another study showed that the number of co-investors an investor had, predicted the likelihood of its portfolio startups raising additional rounds successfully [51].

Media Coverage Media coverage provides legitimacy and credibility to startups. Media attention for startups has been shown to affect the perceived val-

uation of well-informed experts like venture capitalists [**?**]. This has also been shown to translate to increased investment success [8]. 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 TODO

A.1.2.2 Historical Performance

TODO

Financial Performance TODO

Non-Financial Performance TODO

A.1.2.3 Contextual Cues

TODO

Competitor Performance TODO

Local Economy TODO

Broader Economy TODO

APPENDIX B

Original Honours Proposal

Component: Research Proposal

Supervisors: Professor Melinda Hodkiewicz, Dr Tim French

Degree: BPhil(Hons) (24 point project)

University: The University of Western Australia

Background

High-growth technology companies (startups) are turning away from the public markets. Amazon went public in 1997, just two years after its first round of institutional financing, at a market capitalisation of $440M [**?**]. Contrast that with Uber, which remains private six years on and recently raised $3.5B at a $59B pre-money valuation [**?**]. Time to Initial Public Offering (IPO) for Venture Capital (VC)-backed startups has more than doubled over the past 20 years while VC-backed startups pursuing an IPO has plummeted [13].

One explanation for why startups are staying private for longer is the accelerating nature of global business. Startups, particularly those backed by VC firms, are expected to scale fast and require frequent rounds of fundraising coupled with centralized, quick decision making. Such flexibility is not afforded to public companies, due to strict reporting and compliance requirements [52].

Why does this waiting game matter? Principally, because it shifts value creation to the private markets. To put things in perspective, Microsofts market capitalisation grew 500-fold following its IPO [**?**], but for Facebook to do the same now its valuation would have to exceed the global equity market [**?**]. VC funding for late-stage startups is approaching all-time highs, possibly because more investors are entering the private markets to seek higher returns [13].

Merger and Acquisitions (M&A) have far surpassed IPOs as the most common liquidity event for startup founders and investors. In 2015, five times as many US-based VC-backed startups were acquired compared to those that went public through an IPO [13]. Accordingly, startup founders and investors may be interested in predicting which startups are likely to be acquired and by whom. However, M&A prediction is a challenging task.

Previous work has relied on relatively small data sets [50] because publicly-available information on private companies is scarce. In addition, previous work has focused on the financial or managerial features of potential targets [**?**] with little work on textual or social network features.

Xiang and colleagues [**?**] addressed some of these challenges by mining CrunchBase profiles and TechCrunch news articles to predict the acquisition of private startups. Their corpus was larger than previous studies: 38,617 TechCrunch news articles from June 2005 - December 2011 mentioning 5,075 companies, and a total of 59,631 CrunchBase profiles collected in January 2012. Their approach achieved a True Positive rate of between 60-79.8% and a False Positive rate of between 0-8.3%.

There are limitations to Xiang and colleagues' study: the CrunchBase corpus they studied was sparse, only a few common binary classification techniques were tested, and their approach didn't consider IPOs or bankruptcies as potential outcomes. In addition, it is unclear how robust their classifiers are through time. The study could be extended by applying the topic modelling approach to other text corpora such as patent filings, or by attempting a social network link prediction model.

Aim

We aim to produce a supervised learning model that will accurately predict the acquisition of startups in the private markets. We will build on the study by Xiang and colleagues (2012) [**?**], introducing new features and classification techniques. In the previous study, True Positive rate (TP), False Positive rate (FP) and Area under the ROC curve (AUC) were the main evaluation metrics used (collectively, known as "accuracy").

Hypothesis 1 (H1) Xiang and colleagues (2012) [**?**] results can be replicated

H2 Introducing new classification techniques improves accuracy

Xiang and colleagues' study tested three common binary classification techniques: Bayesian Networks (BN), Support Vector Machines (SVM), and Logistic Regression (LR). BN significantly outperformed SVM and LR. The authors suggested that this was because of the high correlation among their features and absence of a linear separator in the feature space. We will test a number of new classification techniques including Random Forests (RF), CART Decision Trees (CART), and Restricted Bolzmann Machines (RBM), to try to improve the accuracy of the model.

H3 Introducing additional CrunchBase features improves accuracy

Xiang and colleagues' study used a total of 22 factual features from CrunchBase profiles. No feature selection process was documented. A recent similar study on AngelList (which has a sharing agreement with CrunchBase) used 85 features of which 11 were selected [8]. Of those 11 features, many were not included in Xiang and colleagues' model. It is plausible that broadening the feature space may result in an improved model.

H4 Introducing additional labels improves accuracy

Xiang and colleagues' study labelled startups as either "acquired" or "not acquired". The "not acquired" category thus includes startups that have bankrupted as well as highly successful startups that went public through an IPO. It is plausible that the breadth of this category would lead to misclassification. Introducing labels for "public" and "bankrupt" could improve the accuracy of the model.

H5 Using more recent CrunchBase corpora improves accuracy

Xiang and colleagues' study used a CrunchBase corpus from January 2012. They found the corpus relatively sparse at the time. Since 2012, the CrunchBase corpus has significantly grown. The CrunchBase Venture Program and the AngelList - CrunchBase data sharing agreement have contributed to the corpus, in addition to natural growth over time. It is plausible that a more recent CrunchBase corpus will provide a better basis for a more accurate model.

This study will improve our understanding of the determinants of startup acquisition in the private markets. The system devised by this study also has the potential to de-risk venture capital and encourage greater investment in private startups.

Method

1. Replicate study by Xiang et al. (2012) [**?**]

   We have requested access to the CrunchBase and TechCrunch datasets used in the previous study (Note: These datasets are currently available on the Carnegie Mellon University intranet). If we are unable to access these datasets we will use a CrunchBase database snapshot from December 2013.

   - Features:
     - Factual Features (CrunchBase)
       - Basic Features e.g. office location, company age
       - Financial Features e.g. investment per funding round
       - Managerial Features e.g. number of acquired companies by founders
     - Topic Features (TechCrunch articles)
   - Outcome: Acquired? (CrunchBase)
   - Processing:
     - Topic model - Latent Dirichlet Allocation (LDA)
     - Classification techniques
       - Bayesian Network (BN)
       - Support Vector Machines (SVM)
       - Logistic Regression (LR)

2. Test additional classification techniques

   - CART Decision Tree (CART) as in [8]
   - Restricted Bolzmann Machine (RBM) as in [8]
   - Random Forest (RF)
   - And other classification techniques

3. Expand the factual features set

   - Founder education (CrunchBase, Dec-2013) as in [8]
   - Founder employment (CrunchBase, Dec-2013) as in [8]
   - Founding team (CrunchBase, Dec-2013) as in [**?**]
   - And other factual features in the CrunchBase corpus

4. Incorporate other potential startup outcomes
    - Outcomes: Bankrupt, Acquired, Public
    - Classification techniques: One vs. all (OVA), All vs. all (AVA)
5. Test classifier robustness over different datasets
    - Original dataset from Xiang et al. (2012) [**?**]
    - CrunchBase readily-available snapshot (December 2013)
    - CrunchBase recent crawl (September 2016)
6. Extend topic modelling and introduce network features (stretch goal)
    - Domain-Constricted LDA model (TechCrunch articles) as in [54]
    - Patent similarity (Google Patents) as in [29]
    - Social network link prediction (CrunchBase) as in [**?**, **?**]
    - And other types of features as time permits

Timeline

Please see below (Table B.1) for a schematic of the proposed methodology.

S:W	**Date**	**Task**
2:03	Fri 19 August	Draft proposal due
2:05	29 Aug - 02 Sep	Proposal defence to research group
2:07	Fri 09 September	Data collected
2:09	Fri 23 September	Replicated previous study
2:SB	Fri 30 September	Draft literature review due
2:12	Fri 28 October	Revised proposal due
2:12	Fri 28 October	Literature review due
2:17	Fri 02 December	Completed main experiments
1:08	Fri 28 April	Draft dissertation due
1:10	Fri 12 May	Seminar title and abstract due
1:13	Mon 29 May	Final dissertation due
1:13	Fri 02 June	Poster due
1:13	29 May - 02 June	Seminar
1:17	Mon 26 June	Corrected dissertation due

Table B.1: Proposed timeline

Software and Hardware Requirements

This project will be developed primarily in Python using scikit-learn, a free open-source machine learning library [**?**]. MySQL may be used to prepare datasets for processing. The system will be hosted on a public compute cloud, likely Amazon Web Services. A free academic license for CrunchBase has been requested.

APPENDIX C

Diagrams

Bibliography

[1] Crunchbase frequently asked questions. `https://info.crunchbase.com/about/faqs/`, 2014. Online; accessed 15 November 2015.

[2] Ahlers, G. K., Cumming, D., Gunther, C., and Schweizer, D. Signaling in equity crowdfunding. *Entrepreneurship Theory and Practice 39*, 4 (2015), 955–980.

[3] Ahn, H., and Kim, K.-j. Using genetic algorithms to optimize nearest neighbors for data mining. *Annals of Operations Research 163*, 1 (2008), 5–18.

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

[5] 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* (2015), AIS Electronic Library.

[6] Arrington, M. The (highly controversial) younoodle startup valuation predictor is coming. `http://techcrunch.com/2008/08/05/the-highly-controversial-younoodle-startup-predictor-is-coming/`, August 2008. Online; accessed 18 May 2015.

[7] 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. *Journal of Business Venturing 19*, 3 (2004), 411–436.

[8] Beckwith, J. Predicting success in equity crowdfunding. Unpublished thesis. Joseph Wharton Research Scholars. Available at `http://repository.upenn.edu/joseph_wharton_scholars/25`, 2016.

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

[10] Blank, S. What's a startup? first principles. https://steveblank.com/2010/01/25/whats-a-startup-first-principles/, 2010. Online; accessed 06 Nov 2016.

[11] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[12] Britz, D., Mai, C., and Xu, C. Quantifying community growth in dynamic social networks. Unpublished thesis. Stanford University. Available at http://snap.stanford.edu/class/cs224w-2013/projects2013/cs224w-007-final.pdf, 2013.

[13] Buyouts Insider. 2016 national venture capital association yearbook. http://www.nvca.org/?ddownload=2963, March 2016. Online; accessed 06 Nov 2016.

[14] 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* (2008), ACM, pp. 96–103.

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

[16] 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* (2016), ACM, pp. 25–30.

[17] Conti, A., Thursby, M., and Rothaermel, F. T. Show me the right stuff: Signals for high-tech startups. *Journal of Economics & Management Strategy 22*, 2 (2013), 341–364.

[18] Cover, T., and Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory 13*, 1 (1967), 21–27.

[19] Croce, A., Guerini, M., and Ughetto, E. Angel financing and the performance of high-tech start-ups. *Journal of Small Business Management* (2016).

[20] Croce, A., Martí, J., and Murtinu, S. The impact of venture capital on the productivity growth of european entrepreneurial firms: 'screening' or 'value added' effect? *Journal of Business Venturing 28*, 4 (2013), 489–510.

[21] Dixon, M., and Chong, J. A bayesian approach to ranking private companies based on predictive indicators. *AI Communications 27*, 2 (2014), 173–188.

[22] Formsma, O. *Entrepreneurs online social networks: Structure and characteristics*. PhD thesis, University of Amsterdam, 2012.

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

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

[25] Geron, T. Real-time venture capital investing: Chi-hua chien's data-based approach. `http://www.forbes.com/sites/tomiogeron/2013/05/08/real-time-venture-capital-investing-chi-hua-chiens-data-based-approach/`, May 2013. Online; accessed 18 May 2015.

[26] Gimmon, E., and Levie, J. Founder's human capital, external investment, and the survival of new high-technology ventures. *Research Policy 39*, 9 (2010), 1214–1226.

[27] Hall, R. E., and Woodward, S. E. The burden of the nondiversifiable risk of entrepreneurship. *The American Economic Review 100*, 3 (2010), 1163–1194.

[28] Hoenen, S., Kolympiris, C., Schoenmakers, W., and Kalaitzandonakes, N. The diminishing signaling value of patents between early rounds of venture capital financing. *Research Policy 43*, 6 (2014), 956–989.

[29] Huang, J., and Zhan, S. With a little help of my (former) employer: Past employment and entrepreneurs' external financing. In *Academy of Management Proceedings* (2015), vol. 2015, Academy of Management, p. 12050.

[30] Hughes, G. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory 14*, 1 (1968), 55–63.

[31] Kotsiantis, S. Supervised machine learning: A review of classification techniques. *Informatica 31*, 3 (2007).

[32] Liang, Y. E., and Yuan, S.-T. D. Predicting investor funding behavior using crunchbase social network features. *Internet Research 26*, 1 (2016), 74–100.

[33] McMullen, J. S., and Dimov, D. Time and the entrepreneurial journey: The problems and promise of studying entrepreneurship as a process. *Journal of Management Studies 50*, 8 (2013), 1481–1512.

[34] McNamara, P. If you had bought 100 shares of microsoft 25 years ago ... `http://www.networkworld.com/article/2228727/data-center/data-center-if-you-had-bought-100-shares-of-microsoft-25-years-ago.html`, March 2011. Online; accessed 06 Nov 2016.

[35] Mingers, J. An empirical comparison of selection measures for decision-tree induction. *Machine learning 3*, 4 (1989), 319–342.

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

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

[38] Patil, A. Crunchbase's venture program members are making startup data better than ever. `https://info.crunchbase.com/2015/01/crunchbases-venture-program-members-are-making-startup-data-better-than-ever/`, January 2015. Online; accessed 18 May 2015.

[39] Raice, S., Das, A., and Letzing, J. Facebook prices ipo at record value. `http://www.wsj.com/articles/SB10001424052702303448404577409923406193162`, May 2012. Online; accessed 06 Nov 2016.

[40] Rish, I. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (2001), vol. 3, IBM New York, pp. 41–46.

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

[42] 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.

[43] Shane, S., and Cable, D. Network ties, reputation, and the financing of new ventures. *Management Science 48*, 3 (2002), 364–381.

[44] Shin, K.-S., Lee, T. S., and Kim, H.-j. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications 28*, 1 (2005), 127–135.

[45] Song, Y., and Vinig, T. Entrepreneur online social networks–structure, diversity and impact on start-up survival. *International Journal of Organisational Design and Engineering 2*, 2 (2012), 189–203.

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

[47] Trachtenberg, A. Changes to our developer program. `https://developer.linkedin.com/blog/posts/2015/developer-program-changes`, February 2015. Online; accessed 18 May 2015.

[48] Tweney, D. What it takes to be a tech entrepreneur in 2015. `http://venturebeat.com/2015/04/30/what-it-takes-to-be-a-tech-entrepreneur-in-2015/`, April 2015. Online; accessed 18 May 2015.

[49] Wang, Z., Zhou, Y., Tang, J., and Luo, J.-D. The prediction of venture capital co-investment based on structural balance theory. *IEEE Transactions on Knowledge and Data Engineering 28*, 2 (2016), 537–550.

[50] Wei, C.-P., Jiang, Y.-S., and Yang, C.-S. Patent analysis for supporting merger and acquisition (m&a) prediction: A data mining approach. In *Workshop on E-Business* (2008), Springer, pp. 187–200.

[51] Werth, J. C., and Boeert, P. Co-investment networks of business angels and the performance of their start-up investments. *International Journal of Entrepreneurial Venturing 5*, 3 (2013), 240–256.

[52] Wies, S., and Moorman, C. Going public: how stock market listing changes firm innovation behavior. *Journal of Marketing Research 52*, 5 (2015), 694–709.

[53] 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).

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

[55] Zhao, X., Zhang, W., and Wang, J. Risk-hedged venture capital investment recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems* (2015), ACM, pp. 75–82.