COVERSHEET

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Faculty of Engineering, Computing and Mathematics

Assignment, Report & Laboratory Coversheet for Individual & Group Assignment

SUBMITTING STUDENT			
SURNAME SHELTON	**GIVEN NAMES** MARK ROBERT	**STUDENT NUMBER** 21151978	
UNIT NAME HONOURS THESIS		**UNIT CODE** CITS4001	
TITLE/TOPIC OF ASSIGNMENT LITERATURE REVIEW		**NAME OF LECTURER/TUTOR** WEI LIU	
DATE/TIME DUE 09-NOV-2016		**DATE/TIME SUBMITTED** 09-NOV-2016	

HONOURS STUDENTS ONLY	OFFICE USE ONLY
By signing this document, I further assert that the length (word count) of my dissertation is within the maximum allowed length governed by the project unit I am enrolled in. Penalties, as outlined on this website, will be applied for over length dissertations.	

FOR GROUP ASSIGNMENTS ONLY	STUDENT NUMBER
NAME	
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

Unless other arrangements have been made it will be assumed that all group members have contributed equally to group assignments/laboratory reports

DECLARATION

I/We are aware of the University's policy on academic conduct (see over) and I/We declare that this assignment/project is my own/my group's work entirely and that suitable acknowledgement has been made for any sources of information used in preparing it. I/We have retained a hard copy for my/our own records.

SIGN:	**SIGN:**
SIGN:	**SIGN:**
SIGN:	**SIGN:**
SIGN:	**SIGN:**

NOTE: No assignment will be accepted without the declaration above being signed and dated

SEE OVER FOR INFORMATION ON REFERENCING & PLAGIARISM

REFERENCING

Information on appropriate referencing (citation) styles can be found under “Manage Your References” at: www.is.uwa.edu.au/information-resources/guides

Note that:
Each drawing, picture, photograph, quotation or block of text copied from a source must be acknowledged individually. This can be done using a referencing style (see above) or by including the full reference in the text or in a footnote. It is not sufficient to simply list sources in a bibliography at the end without including the individual references to the sources in the main text.

The same rules apply to materials taken from the web. The authorship and source must be traceable.
The boundaries between your original work and copied work must be clear. Use quotes, indentation and/or font style to make the distinction clear.

PLAGIARISM

“The appropriation or imitation of another’s ideas and manner of expressing them to be passed off as one’s own”.

(The Macquarie Dictionary, 1981)

Synonyms:

Piracy, copying, forgery, lifting, expropriation, appropriation

“Plagiarism is the unattributed use of someone else's words, creations, ideas and arguments as one's own. Within university policies it is usually further extended to include the use of 'too close' or extensive paraphrasing. For example, cutting and pasting text from the Web without attributing it to the author is plagiarism and therefore dealt with as cheating. Similarly, substituting a few words of copied text without changing the structure of the piece also constitutes plagiarism. There is a range of penalties for academic misconduct, depending on the seriousness of the cheating, from loss of credit to expulsion from the University.” (UWA Handbooks 2013)

The University of Western Australia treats plagiarism as serious academic misconduct. The University can impose severe penalties, including expulsion. Refer to Statue 17 Student Discipline and the associated Regulations for Student Conduct and Discipline at www.uwa.edu.au/current/information/discipline

See also

Faculty Policy on Plagiarism:

www.ecm.uwa.edu.au/students/exams/dishonesty

UWA’s policy statement on Ethical Scholarship, Academic Literacy and Academic Misconduct: www.handbooks.uwa.edu.au/postgraduate/policies

Factors That Influence Startup Investment

Literature Review

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

Contents

CHAPTER 1

Literature Review

Technological advances have made launching startups more accessible than ever before. Customers are easily accessed through the Internet and launching a startup can be done from a bedroom. However, startups remain competitive and risky endeavours. Startups can be unprofitable for years so entrepreneurs look for incubators, accelerators, angel investors and venture capital firms to support them through this developmental period. Aside from funding, investors hold experience and networks that can accelerate startup growth. Investors act as scouts, able to identify the potential of new startups, and as coaches, able to help startups realise that potential [5].

Startups must convince investors to support them throughout their development, but this process can be burdensome and time-consuming. Investors find it difficult to evaluate startup potential for investment because metrics of performance often do not exist or are uncertain [44]. The popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of a desire for better methods of assessing startup potential. By 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 [36].

Investment comes with trade-offs for startups. The majority of venture capital-backed startups end in bankruptcy [41]. Investors are protected from these losses because the minority of their investments that are successful have outsized returns: 85% of venture capital returns come from 10% of investments [41]. Investors seek to optimise the risk-reward trade-off by pressing startups to grow rapidly, frequently raise funding rounds and make quick, centralised decisions [16]. The rapid growth demanded by investors is generally incompatible with public company structures, due to reporting and compliance requirements [55]. Accordingly, we observe venture capital-backed startups delaying Initial Public Offerings (IPO). Time taken to IPO has doubled in the past 20 years [34].

Startups remaining privately-held for longer shifts value creation to the pri-

vate markets. Microsoft's market capitalisation grew 500-fold following its IPO in 1986, but for Facebook to grow to the same extent since its IPO in 2012 its capitalisation would exceed the total global equity market. Investment in late-stage startups is approaching all-time highs as public market investors enter the private markets [34]. Given this situation, it is important to understand how factors that influence investment change through a startup's development. A clear gap in the academic literature exists in this area. Studies of the determinants of startup investment have common weaknesses. This study will address these weaknesses in three ways:

Larger Sample Size Previous studies are restricted in sample size. Most studies have samples of fewer than 500 startups [1, 20], or between 500 and 2,000 startups [26, 58, 3, 54, 14], and only a few have large scale samples (more than 100,000 startups) [43, 12]. Sample size is more critical to model development than the sophistication of machine learning algorithms or feature selection [10]. 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.

Developmental Focus Prior work focuses on early-stage investment in startups, primarily equity crowdfunding [6, 1, 12, 59] and angel investing [14]. The functions and objectives of startups change through their development [32]. For example, early stages of funding are characterised by uncertainty about technical validity and market fit [27]. In this setting, patents are a strong signal to investors, but may become less so in later rounds. Generally, we expect signals that attract investment in startups will change over time.

Rich Features Prior work focuses on basic company features (e.g. the headquarters' location, the age of the company) for startup investment predictive models [6, 20]. Semantic text features (e.g. patents, media) [26, 59] and social network features (e.g. co-investment networks) [54, 12, 58] 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 will develop software that collects and processes information about startups to predict their likelihood of raising investment at different stages of their development. This study has potential for scholarly, policy and firm-specific implications. Our scholarly contribution is a conceptual framework for startup

investment, based on work by Ahlers et al. [1]. Our conceptual framework posits that startup investment is a product of two factors: startup potential and investment confidence. We will test this framework with respect to startup development using cross-sectional and longitudinal analyses. We aim to contribute to the understanding of the determinants of startup investment, with a focus on how they change over time. Ultimately, we hope we can encourage greater investment in startups.

The paper proceeds as follows. The next section explores theoretical models of technology startups and startup investment (Section 1.1). Thereafter, we review empirical evidence of features linked to startup investment (Section 1.2). We then determine how to collect the data to test those features (Section 1.3) and evaluate machine learning algorithms to find those that suit this startup investment prediction task (Section 1.4). The final section summarises our main results and concludes.

1.1 Theoretical Background

Startups are a medium for the commercialisation of new and disruptive technologies or business models. In this sense, startups perform the critical role of Schumpeterian entrepreneurship, or "creative destruction", in the economy [51]. Inherent to this role is high risk of failure [41]. Therefore, it is important for policy-makers, investors and entrepreneurs to understand the factors that drive startup performance. We discuss drivers of entrepreneurial growth and the role of investment, and propose a conceptual framework for startup investment.

1.1.1 Technology Startups

Steve Blank, entrepreneur-mentor and author, defines a startup as "an organization formed to search for a repeatable and scalable business model." [8]. In this case, "search" differentiates established startups from small businesses, such as restaurants. While startups seek to disrupt industries in novel ways, there are many similarities in their underlying mechanisms. We review the prototypical startup development lifecycle and common factors that contribute to startup growth and performance.

1.1.1.1 Startup Lifecycle

The functions and objectives of startups change through their development [32]. The lifecycle of a startup can be divided into stages that reflect these shifts (Figure 1.1). These stages may be rapid or prolonged, depending on the intensity of technological development. First, founders identify a need they wish to solve. Second, founders develop an idea or minimal viable product (MVP) for a product or service that solves their identified need. Startups may raise their first funding round at this stage through friends and family, angel investors or equity crowdfunding. Third, founders create a company structure and develop an organisation that allows their solution to be developed, marketed and sold. Once they have evidence of market adoption, founders may raise funding rounds from venture capital firms to accelerate their growth. Fourth, the startup finds product-market fit and continues to grow, heads toward an IPO or is acquired. Alternatively, and more likely, the startup might fail or revert to address another problem or solution [22].

1.1.1.2 Determinants of Performance

Traditional assessments of company valuation (e.g. discounted cashflows, comparable valuation) are difficult to apply to startups because they change so rapidly. Instead, when valuing startup potential we often seek more fundamental drivers. The performance of large companies is dominated by external factors: macroeconomic and industry trends, and competitor activity [23]. However, startups are small and nimble enough to be largely unaffected by these factors. Rather, key determinants of startup performance are understood to be internal, flexible to changes in product and market. These determinants are often categorised into three fundamental elements: human capital, social capital and structural capital [5, 1].

Human Capital Human capital is value derived from education, experience and skills [20]. Early-stage startups have little financial and structural capital and so rely on human capital to drive company growth. Human capital helps entrepreneurs identify and exploit business opportunities, define and realise their strategies, and acquire additional resources (including funding) [1]. A meta-review of human capital in startups shows industry-related experience, education and founding team compatibility contribute to startup survival [20]. Venture capitalists indicate experience and management skills are key selection criteria for early-stage startup investments [20].

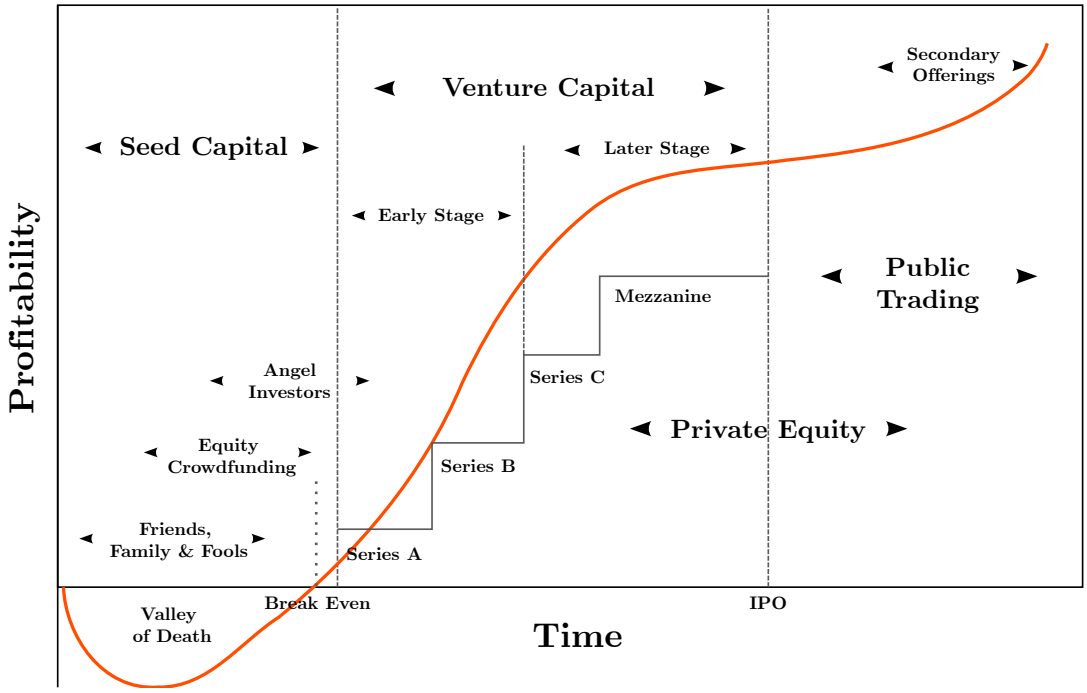

Figure 1.1: Idealised startup development lifecycle (adapted, [56]). Red line represents profitability over time. Common financial instruments are labelled at times they are most likely to coincide. The chart is divided into three periods: an early period of unprofitability ("Valley of Death") where seed capital supports the business, a period of growth sustained by rounds of venture capital, and a transition to stability and mature capital markets.

Social Capital Social capital is value derived from relationships [19]. Startups require resources such as advice, finance, skills and labour to be able to realise entrepreneurial opportunities [5]. Social networks provide the media for those resources to be obtained. Social capital is a determinant of entrepreneurial performance, including survival time [47], venture capital raised [21] and revenue generated [48]. Entrepreneurs centrally embedded in social networks are more likely to access necessary resources [47]. Being an influencer or aligning oneself with influencers can increase the quality of an entrepreneur's connections [21]. Strengthening and maintaining social networks plays an essential role in startup performance.

Structural Capital Structural capital is value derived from intangible assets, infrastructure, and systems. Intellectual property and their proxy, patents, are a key component of structural capital, especially for early-stage startups. Innovation is a key determinant of firm survival. Patents support

the appropriation of returns from innovative activities and facilitate cooperation and bargaining with business partners. Indeed, patent ownership is correlated with startup valuation [27], and leads to greater performance in terms of asset growth [24] and likelihood of survival [53]. Similarly, studies demonstrate a positive relationship between patent filings and startup investment [26].

1.1.2 Startup Investment

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

1.1.2.1 Benefits of Investment

Many obstacles confront young companies. Startups operate using immature routines, with little knowledge of their environment and poor working relationships with customers and suppliers. In addition, startups require substantial resources to fund speculative development projects, especially when commercialising new technologies [18]. For these reasons, venture capital is highly desirable for entrepreneurs. Venture capital-backed firms grow faster, patent more, have higher productivity and are more likely to go public [5, 25]. This effect can be explained in three ways: first, investors are effective scouts of successful startups; second, investors directly help startups become successful; and third, investors send a signal that encourages other third parties (e.g. future employees, customers, investors) to support the startup.

1.1.2.2 Investment Techniques

Venture capital firms manage and invest third-party funds. Venture capital firms seek investments capable of providing financial returns and a successful exit (IPO or acquisition) within the required time frame of their fund (37 years). Accordingly, venture capital firms are highly selective and spend much of their time searching for potentially attractive startups and evaluating their potential for investment. There are two common approaches investors use to evaluate startup potential. First, investors extrapolate current performance metrics (e.g. app

downloads, viral momentum, lifetime value of customers). Second, investors evaluate determinants of startup performance (e.g. human capital, socical capital). Regardless of approach, a key challenge for investors is informational asymmetry. Founders possess more information than investors about a startup's potential and expecting founders to transfer that knowledge to potential investors is unrealistic. To evaluate the quality of a startup's signals, investors look to factors such as third party validation (e.g. previous investments), historical performance (e.g. profitability), and contextual cues (e.g. competitor performance) [1, 26, 27].

1.1.3 Proposed Framework

External investment is a key driver of startup development. However, our understanding of factors that influence startup investment is incomplete. We discuss two key factors in investment decisions: first, evaluating startup potential and second, gauging confidence in that evaluation. Ahlers et al. [1] develop 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 1.2.

1.2 Feature Selection

We develop a conceptual framework relating startup potential and investor confidence to startup investment. We seek to operationalise this conceptual framework into features 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 A, we describe each of these features and outline conceptual and empirical evidence that justify their inclusion in our conceptual framework.

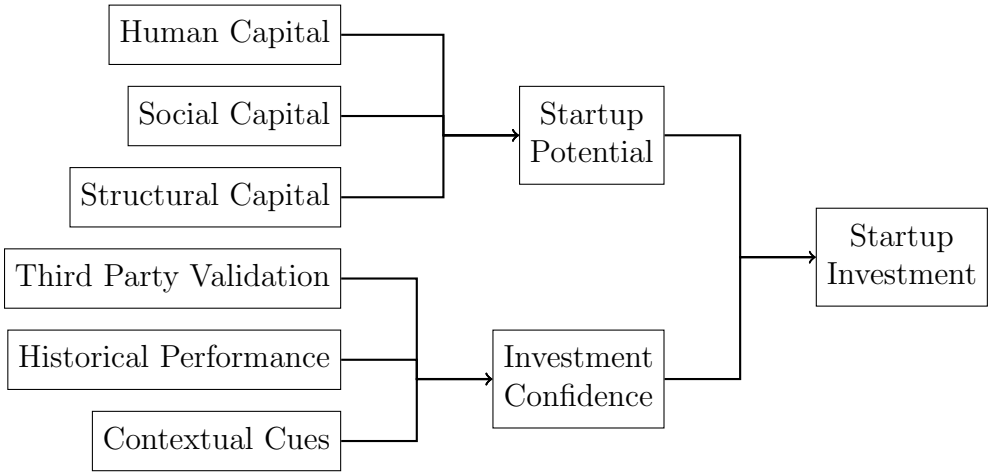

Figure 1.2: Proposed conceptual framework for startup investment. We adapt the framework proposed by Ahlers et al. [1], originally based on work by Baum and Silverman [5]. For an extended version of this framework, please refer to Figure A.1.

Feature selection is critical to the success of our proposed conceptual framework. We build 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 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 1.2, we outline the characteristics of relevant data sources and how they could contribute to our chosen

Features	Results from Studies	
	Significant	Non-Significant
Startup Potential		
Human Capital		
Founders' Capabilities	[6, 3, 20]	[43, 13]
NED Capabilities	[5]	[1, 3]
Staff Capabilities	[6, 3, 13]	[1]
Social Capital		
Strategic Alliances	[5]	-
Social Influence	[6, 3, 12, 58]	-
Structural Capital		
Patent Filings	[26, 27, 5]	[1, 20]
Investment Confidence		
Third Party Validation		
Investment Record	[1, 6, 14, 26, 13]	-
Investor Reputation	[3, 54, 27]	[26]
Media Coverage	[6]	[3]
Awards and Grants	[1]	-
Historical Performance		
Financial Performance	[6, 5]	-
Non-Financial Performance	[3, 20]	[26]
Contextual Cues		
Competitor Performance	[43, 14, 20]	[6, 13]
Broader Economy	[6, 14, 26, 13, 27]	[43, 1]
Local Economy	[43, 6, 14, 20, 26]	-

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.

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.

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.

1.3.1 Source Characteristics

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 will only consider online data sources for inclusion in this study. In addition, we will restrict our scope to technology companies based in the United States. This sample is well-represented in the data sources we review. Next, we review the characteristics of online data sources commonly used in entrepreneurship research.

1.3.1.1 Startup Databases

Startup databases collect and store information about startups, investors, media coverage 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. 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.

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 active online community to contribute to and edit most of its pages. However, this results in unpopular startups having relatively sparse profiles. CrunchBase has three provisions to prevent and remediate inaccurate crowd-sourced entries. First, users authenticate their 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. Third, established

startups have their editing privileges locked and updates require manual verification.

AngelList AngelList combines 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 [12]. AngelList 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, which is useful for longitudinal studies.

1.3.1.2 Social Media

Social media platforms allow people to network and communicate online. They also capture information about peoples' identities and relationships that can be used in research. Social networking sites are diverse. Two social networks studied in detail in entrepreneurship research are LinkedIn and Twitter.

LinkedIn LinkedIn is a social network used for professional networking. It is commonly used in social network studies of entrepreneurship because it holds human capital information including past employment and education [47]. 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 [52], making it difficult to use for social network analyses.

Twitter Twitter is a social networking and micro-blogging site often used by entrepreneurs to promote their personal and business brands and share news and opportunities [47]. 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 descriptions). However, Twitter's API only provides Tweets published within the last 7

days and access to historical Twitter data requires a commercial license [38].

1.3.1.3 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 [42]. 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).

Financial Reports Acquiring private company 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. 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 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].

1.3.2 Source Evaluation

We evaluate 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 and timestamping. At least one startup database should be used and either are satisfactory. Of the other data sources we review, patent filings and Twitter are most promising. PatentsView provides comprehensive patent information, though it could prove difficult matching identities to other sources. Twitter provides social network topology and basic profile information through its free API but does not provide access to historical tweets. Other data sources are less promising because of access issues. LinkedIn cannot

be easily collected now the API is deprecated. Financial reports are too expensive for the purposes of this study.

1.4 Learning Algorithms

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." [33].

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 evaluate 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 based on whether they receive investment at each funding round (Seed, Series A, B etc.). We may also investigate measures of startup performance (e.g. survival time, exit). 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

Criteria	Machine Learning Algorithms													
	NB		LR		KNN		DT		RF		SVM		ANN	
Data Set Properties	**2**		4		6		**2**		**1**		5		7	
Missing Values	✓✓	[29]	✓	-	✗	[29]	✓✓	[29]	✓✓	[49]	✓	[29]	✗	[29]
Mixed Feature Types	✓✓	[29]	✓✓	-	✓✓	[29]	✓✓	[29]	✓	[49]	✓	[29]	✓	[29]
Irrelevant Features	✗	[29]	✗	[30]	✓	[29]	✓✓	[29]	✓✓	[49]	✗	[29]	✗	[29]
Imbalanced Classes	✓✓	-	✓✓	-	✗	-	✗	[29]	✓	[49]	✓✓	[29]	✓	[29]
Algorithm Properties	**2**		**1**		5		5		**2**		5		**2**	
Predictive Power	✗	[10]	✓	[10]	✓	[10]	✗	[29]	✓✓	[10]	✓✓	[10]	✗✓	[10]
Interpretability	✓✓	[29]	✓✓	[30]	✗	[29]	✓✓	[29]	✓	[30]	✗	[29]	✗	[29]
Incremental Learning	✓✓	[29]	✓✓	-	✓✓	[29]	✓	[29]	✓	-	✓	[29]	✓✓	[29]
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.

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

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 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.

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 feature types is to convert into a unified type (e.g. real-valued, binary). However, this process partially discards type-specific information. We expect mixed feature types in our data set as we will be handling data from databases, social networks and semantic text analysis. Therefore, robustness to mixed feature types is a desirable property of our algorithm.

Irrelevant Features 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 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.

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 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

Predictive Power 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, then 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 too.

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 a key objective that we improve our understanding of the determinants of startup investment. Therefore, interpretability is a desirable property of our algorithm.

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 change over time. Incremental learning will allow us to gain an understanding of whether that is accurate. Therefore, incremental learning is a desirable property of our algorithm.

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 the following sections, we describe each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in applications relevant to startup investment.

1.4.2.1 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 [29]. However, Naive Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features [35]. 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.

1.4.2.2 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 [17], 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 [10]. 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 [20].

1.4.2.3 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 [29]. 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.

1.4.2.4 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 [29]. 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 [11]. 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 [31]. Decision Trees had the highest accuracy and are desirable because their reasoning is easily communicated to startups.

1.4.2.5 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 [9]. 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 [11]. 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 [9].

1.4.2.6 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 high predictive power [11], 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 [46]. Support Vector Machines outperformed Artificial Neural Networks, possibly because of the small data set.

1.4.2.7 Artificial Neural Networks

Artificial Neural Networks are a computational approach based on a network of neural units (neurons) that loosely models the way 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 [40]. 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.

1.4.3 Algorithm Evaluation

We evaluate 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 [10]. 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 Conclusion

We conduct a literature review to determine the factors that influence investment for startups. First, we explore theoretical models of technology startups

and startup investment (Section 1.1). Thereafter, we reviewe empirical evidence of features linked to startup investment (Section 1.2). We determine how to collect the data to test those features (Section 1.3) and evaluate machine learning algorithms to find those that suit startup investment prediction (Section 1.4).

Venture capital funding for late-stage privately-held startups is approaching all-time highs as investors enter the private markets [34]. 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 predicting startup investment. Existing approaches in the literature have three common limitations: small sample size [1, 20, 15, 26, 58, 3, 54, 14], a focus on early-stage investment [6, 1, 12, 59, 14], and sparse use of features [1, 3, 12, 14, 54, 20]. Although individual studies address some of these limitations, none attempt to synthesise their findings into a standalone study and software design.

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 et al. [1] and Baum and Silverman [5]. Our conceptual framework models startup investment as a product of two factors: startup potential and investment confidence. We will test this framework with respect to startup development, using cross-sectional and longitudinal analyses. Our aim is that this study improves our understanding of the determinants of startup investment, especially in later-stage startups. Ultimately, we hope this study encourages greater investment in startups.

APPENDIX A

Feature Summary

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, 20], as is whether a founder has obtained an MBA [6]. In addition, past entrepreneurial experience seems to be a predictive factor [20] though there is some evidence to dispute this [43]. Finally, the number of founders seems to be correlated

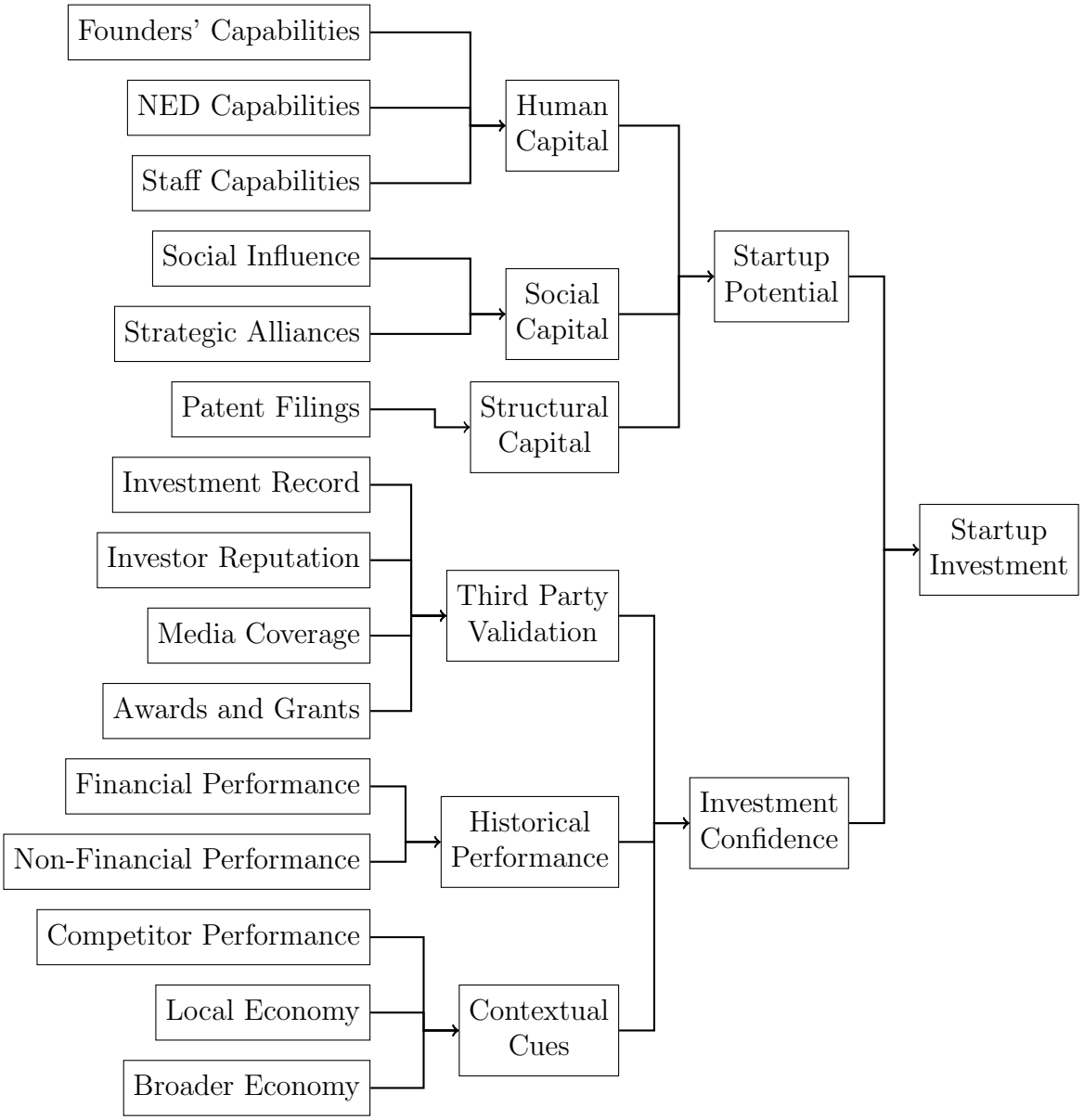

Figure A.1: Proposed conceptual framework for startup investment. This extended version of the framework includes features identified by empirical studies of startup investment. We adapt the framework proposed by Ahlers et al. [1], originally based on work by Baum and Silverman [5].

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 [28]. 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 [28]. 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, 13]. 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 [45]. 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 [12, 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 [50]. 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, 26]. However, there may not be as strong a relationship for non-biotechnology startups (e.g. software) [20, 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 [27, 25].

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, 14, 26, 13].

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 [25]. Followers on AngelList and previous co-investors predict the likelihood of an investor's portfolio startups raising additional rounds successfully [3, 54].

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 [37]. 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 [57], however, most studies are only able to utilise simplistic performance metrics like survival time [39, 47, 21].

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 [47, 3, 20]. 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 [41].

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 [43, 14, 20].

Local Economy Headquarters location is a key indicator of startup investment success [6, 14, 20]. 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 [14, 26].

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] Dixon, M. and Chong, J. "A Bayesian approach to ranking private companies based on predictive indicators". In: *AI Communications* 27.2 (2014), pp. 173–188.

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

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

[18] Gans, J. S. and Stern, S. "The product market and the market for "ideas": commercialization strategies for technology entrepreneurs". In: *Research policy* 32.2 (2003), pp. 333–350.

[19] Gedajlovic, E., Honig, B., Moore, C. B., Payne, G. T., and Wright, M. "Social capital and entrepreneurship: A schema and research agenda". In: *Entrepreneurship Theory and Practice* 37.3 (2013), pp. 455–478.

[20] 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.

[21] 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.

[22] Hall, R. E. and Woodward, S. E. "The burden of the nondiversifiable risk of entrepreneurship". In: *The American Economic Review* 100.3 (2010), pp. 1163–1194.

[23] Hansen, G. S. and Wernerfelt, B. "Determinants of firm performance: The relative importance of economic and organizational factors". In: *Strategic Management Journal* 10.5 (1989), pp. 399–411.

[24] Helmers, C. and Rogers, M. "Does patenting help high-tech start-ups?" In: *Research Policy* 40.7 (2011), pp. 1016–1027.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] McMullen, J. S. and Dimov, D. "Time and the entrepreneurial journey: The problems and promise of studying entrepreneurship as a process". In: *Journal of Management Studies* 50.8 (2013), pp. 1481–1512.

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

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

[35] 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.

[36] Patil, A. *CrunchBase's Venture Program Members Are Making Startup Data Better Than Ever*. Ed. by Crunchbase.com. `https://info.crunchbase.com/2015/01/crunchbases-venture-program-members-are-making-startup-data-better-than-ever/`. Online; accessed 18 05 2015. Jan. 2015.

[37] 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.

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

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

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

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

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

[43] 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.

[44] Shane, S. and Cable, D. "Network ties, reputation, and the financing of new ventures". In: *Management Science* 48.3 (2002), pp. 364–381.

[45] 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.

[46] 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.

[47] 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.

[48] Stam, W. and Elfring, T. "Entrepreneurial orientation and new venture performance: The moderating role of intra-and extraindustry social capital". In: *Academy of Management Journal* 51.1 (2008), pp. 97–111.

[49] 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.

[50] 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.

[51] Timmons, J. A. and Bygrave, W. D. "Venture capital's role in financing innovation for economic growth". In: *Journal of Business Venturing* 1.2 (1986), pp. 161–176.

[52] 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.

[53] Wagner, S. and Cockburn, I. "Patents and the survival of Internet-related IPOs". In: *Research Policy* 39.2 (2010), pp. 214–228.

[54] 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.

[55] Wies, S. and Moorman, C. "Going public: how stock market listing changes firm innovation behavior". In: *Journal of Marketing Research* 52.5 (2015), pp. 694–709.

[56] Wikimedia Commons. *Diagram of the typical financing cycle for a startup company*. `"https://commons.wikimedia.org/wiki/File:Startup_financing_cycle.svg"`. Online; accessed 07 Nov 2016. Mar. 2009.

[57] 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.

[58] 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.

[59] 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.

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 [5]. Contrast that with Uber, which remains private six years on and recently raised $3.5B at a $59B pre-money valuation [2]. 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 [7].

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 [13].

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 [6], but for Facebook to do the same now its valuation would have to exceed the global equity market [9]. VC funding for late-stage startups is approaching all-time highs, possibly because more investors are entering the private markets to seek higher returns [7].

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 [7]. 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 [12] because publicly-available information on private companies is scarce. In addition, previous work has focused on the financial or managerial features of potential targets [3] with little work on textual or social network features.

Xiang and colleagues [14] 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) [14], 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) [14] 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 [1]. 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) [14]

   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 [1]
   - Restricted Bolzmann Machine (RBM) as in [1]
   - Random Forest (RF)
   - And other classification techniques

3. Expand the factual features set

   - Founder education (CrunchBase, Dec-2013) as in [1]
   - Founder employment (CrunchBase, Dec-2013) as in [1]
   - Founding team (CrunchBase, Dec-2013) as in [11]
   - 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) [14]
    - 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 [15]
    - Patent similarity (Google Patents) as in [4]
    - Social network link prediction (CrunchBase) as in [10, 16]
    - 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 [8]. 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.

Bibliography

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

[2] Buhr, S. *Uber takes its most significant investment yet at $3.5 billion from Saudi Arabia*. https://techcrunch.com/2016/06/01/uber-takes-its-most-significant-investment-yet-at-3-5-billion-from-saudi-arabia/. Online; accessed 07 Nov 2016. June 2016.

[3] Hongjiu, L., Huimin, C., and Yanrong, H. "Financial characteristics and prediction on targets of M&A based on SOM-Hopfield neural network". In: *2007 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE. 2007, pp. 80–84.

[4] Huang, J. and Zhan, S. "With a Little Help of My (Former) Employer: Past Employment and Entrepreneurs' External Financing". In: *Academy of Management Proceedings*. Vol. 2015. 1. Academy of Management. 2015, p. 12050.

[5] Kawamoto, D. *Amazon.com IPO skyrockets*. http://www.cnet.com/au/news/amazon-com-ipo-skyrockets/. Online; accessed 07 Nov 2016. May 1997.

[6] McNamara, P. *If you had bought 100 shares of Microsoft 25 years ago ...* Ed. by Network World. http://www.networkworld.com/article/2228727/data-center/data-center-if-you-had-bought-100-shares-of-microsoft-25-years-ago.html. Online; accessed 06 Nov 2016. Mar. 2011.

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

[8] Pedregosa, F. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* (2011).

[9] Raice, S., Das, A., and Letzing, J. *Facebook prices IPO at record value*. Ed. by Journal, T. W. S. http://www.wsj.com/articles/SB10001424052702303448404577409
Online; accessed 06 Nov 2016. May 2012.

[10] Shi, Z., Lee, G. M., and Whinston, A. B. "Towards a better measure of business proximity: topic modeling for analyzing M As". In: *Proceedings of the 15th ACM Conference on Economics and Computation*. ACM. 2014, pp. 565–565.

[11] Spiegel, O., Abbassi, P., Schlagwein, D., and Fischbach, K. "Going it all alone in web entrepreneurship?: a comparison of single founders vs. co-founders". In: *Proceedings of the 2013 Annual Conference on Computers and People Research*. ACM. 2013, pp. 21–32.

[12] 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*. Springer. 2008, pp. 187–200.

[13] Wies, S. and Moorman, C. "Going public: how stock market listing changes firm innovation behavior". In: *Journal of Marketing Research* 52.5 (2015), pp. 694–709.

[14] Xiang, G., Zheng, Z., Wen, M., Hong, J. I., Rose, C. P., and Liu, C. "A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch." In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. AAAI. 2012.

[15] 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.

[16] Yuxian, E. L. and Yuan, S.-T. D. "Investors Are Social Animals: Predicting Investor Behaviour using Social Network Features via Supervised Learning Approach". In: *International Workshop On Mining And Learning With Graphs*. 2013.

APPENDIX C

Revised 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

Technological advances have made launching startups more accessible than ever before. Customers can be accessed easily through the Internet and launching a startup can be done from a bedroom. However, startups remain competitive and risky endeavours. Startups can be unprofitable for years so entrepreneurs look for incubators, accelerators, angel investors and venture capital firms to support them through this developmental period. Aside from funding, investors hold experience and networks that can accelerate startup growth. Investors act as scouts, able to identify the potential of new startups, and as coaches, able to help startups realise that potential [3].

Startups must convince investors to support them throughout their development, but this process can be burdensome and time-consuming. Investors find it difficult to evaluate startup potential for investment because metrics of performance often do not exist or are uncertain [18]. Popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of a desire for better methods of assessing startup potential. By 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 [14].

Investment comes with trade-offs for startups. The majority of venture capital-backed startups end in bankruptcy [16]. Investors are protected from these losses

because the minority of their investments that are successful have outsized returns: 85% of venture capital returns come from 10% of investments [16]. Investors seek to optimise the risk-reward trade-off by pressing startups to grow rapidly, frequently raise funding rounds and make quick, centralised decisions [8]. The rapid growth demanded by investors is generally incompatible with public company structures, due to reporting and compliance requirements [20]. Accordingly, we see venture capital-backed startups delaying Initial Public Offerings (IPO). Time taken to IPO has doubled in the past 20 years [13].

Aim

Startups remaining privately-held for longer shifts value creation to the private markets. Microsoft's market capitalisation grew 500-fold following its IPO in 1986, but for Facebook to grow to the same extent since its IPO in 2012 its capitalisation would exceed the total global equity market. Investment in late-stage startups is approaching all-time highs as public market investors enter the private markets [13]. Given this situation, it is important to understand how factors that influence investment change through a startup's development. A clear gap in the academic literature exists in this area. Studies of the determinants of startup investment have common weaknesses. This study will address these weaknesses in three ways:

Larger Sample Size Previous studies are restricted in sample size. Most studies have samples of fewer than 500 startups [1, 9], or between 500 and 2,000 startups [10, 21, 2, 19, 7], and only a few have large scale samples (more than 100,000 startups) [17, 6]. Sample size is more critical to model development than the sophistication of machine learning algorithms or feature selection [5]. 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.

Developmental Focus Prior work focuses on early-stage investment in startups, primarily equity crowdfunding [4, 1, 6, 22] and angel investing [7]. The functions and objectives of startups change through their development [12]. For example, early stages of funding are characterised by uncertainty about technical validity and market fit [11]. In this setting, patents are a strong signal to investors, but may become less so in later rounds. Generally, we expect signals that attract investment in startups will change over time.

Rich Features Prior work focuses on basic company features (e.g. the headquarters' location, the age of the company) for startup investment predictive models [4, 9]. Semantic text features (e.g. patents, media) [10, 22] and social network features (e.g. co-investment networks) [19, 6, 21] 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 will develop software that collects and processes information about startups to predict their likelihood of raising investment at different stages of their development. This study has potential for scholarly, policy and firm-specific implications. Our scholarly contribution is a conceptual framework for startup investment, based on work by Ahlers et al. [1]. Our conceptual framework posits that startup investment is a product of two factors: startup potential and investment confidence. We will test this framework with respect to startup development using cross-sectional and longitudinal analyses. We aim to contribute to the understanding of the determinants of startup investment, with a focus on how they change over time. Ultimately, we hope that we can encourage greater investment in startups.

Method

Data Collection We will develop an automated data collection system that will provide a platform on top of which we can easily perform our analyses. Our primary data sources are CrunchBase, AngelList, Twitter and PatentsView. We will start with a focus on CrunchBase and then develop systems to match the other sources. CrunchBase data can be accessed in multiple ways. The simplest format are comma separated files (CSV) that hold data about each relation in their database (e.g. funding rounds, investors). A current CSV dump of the database can be requested at any time from CrunchBase. We have also retrieved several older CSV dumps that can be compared with current data for longitudinal studies. CSV dumps provide a subset of the attributes in the CrunchBase data set. To get the full data set requires access through their application programming interface (API). We will develop a crawler that can continually traverse the API iteratively to effectively mirror the CrunchBase data set locally for further analyses. Our master database is likely to be a Sqlite server. We are also investigating distributed solutions, including using Spark.

Machine Learning Analyses 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 receives funding at each funding round. We may also investigate measures of startup performance (e.g. survival time, exit). We will compare and evaluate machine learning algorithms to find which algorithms suits this task best. We have collected six historical CSV dumps from CrunchBase spanning the period from October 2013 to the present. We will match companies across these data sets to test the robustness of our model across time and to see whether the gradient of change in different features can provide greater accuracy to our model than the static features.

Timeline

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

S:W	**Date**	**Task**
2:03	Fri 19 August	Draft proposal due
2:14	Wed 09 November	Revised proposal due
2:14	Wed 09 November	Literature review due
2:16	Fri 25 November	Data collected
2:21	Fri 30 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 C.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 [15]. Sqlite may be used to prepare datasets for processing. The system will be hosted on a public compute cloud, likely Amazon Web Services. Free academic licenses for CrunchBase and AngelList have been requested.

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] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] McMullen, J. S. and Dimov, D. "Time and the entrepreneurial journey: The problems and promise of studying entrepreneurship as a process". In: *Journal of Management Studies* 50.8 (2013), pp. 1481–1512.

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

[14] Patil, A. *CrunchBase's Venture Program Members Are Making Startup Data Better Than Ever*. Ed. by Crunchbase.com. `https://info.crunchbase.com/2015/01/crunchbases-venture-program-members-are-making-startup-data-better-than-ever/`. Online; accessed 18 05 2015. Jan. 2015.

[15] Pedregosa, F. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* (2011).

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

[17] 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.

[18] Shane, S. and Cable, D. "Network ties, reputation, and the financing of new ventures". In: *Management Science* 48.3 (2002), pp. 364–381.

[19] 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.

[20] Wies, S. and Moorman, C. "Going public: how stock market listing changes firm innovation behavior". In: *Journal of Marketing Research* 52.5 (2015), pp. 694–709.

[21] 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.

[22] 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.