**(0.) Introduction**

**(0.1) Background**

Technological advances have made launching a startup more accessible than ever before \cite{tweney2015}. Billions of consumers can be accessed through the Internet and launching a startup can be done from a bedroom. However, startups remain competitive and risky endeavours. Technology startups can be unprofitable for years so entrepreneurs look for incubators, accelerators, angel investors and venture capitalists to support them through this developmental period. Aside from funding, investors hold other key resources (e.g. information, networks) that accelerate startup growth \cite{croce2013}. Investors act as scouts, able to identify the potential of new startups \cite{baum2004}, and as coaches, able to help startups realise that potential \cite{croce2013}.

It’s important for startups to continue to win over investors throughout their development, but this is not a trivial task. Raising funding rounds from investors can be challenging and time-consuming, as investors find it difficult to quickly evaluate startups as investment opportunities. Investors spend time seeking and evaluating signals of a startup's underlying quality because clear metrics of performance often do not exist or are difficult to capture \cite{shane2002}. The growing popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of a desire for a more efficient assessment of startup potential. By the end of 2014, over 1200 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 \cite{patil2015}.

Venture capital investment does not come without its trade-offs. About 50% of venture capital-backed startups end in complete liquidation \cite{hall2010}. Venture capital investors are protected from these losses because the minority of their investments that do survive have outsized returns: 85% of venture capital returns come from just 10% of investments \cite{sahlman2010}. As the venture capital industry matures, investors optimise for this extreme risk-reward trade-off by pushing startups to grow rapidly, frequently raise follow-on funding rounds and make quick, centralised decisions \cite{manigart2002}. The rapid growth demanded by venture capital investors is generally incompatible with public company structures, due to strict reporting and compliance requirements \cite{wies2015}. Accordingly, venture-capital backed startups are delaying their Initial Public Offerings (IPO): time to IPO has doubled in the past 20 years \cite{nvca2016}.

**(0.2) Rationale**

Startups remaining privately-held for longer has the effect of shifting value creation to the private markets. For example, Microsoft’s market capitalisation grew 500-fold following its IPO in 1986 \cite{microsoft}, but for Facebook to grow to the same extent since its IPO in 2012 its valuation would exceed the capitalisation of the global equity market \cite{facebook}. Venture capital funding for late-stage privately-held startups is approaching all-time highs as investors enter the private markets \cite{nvca2016}. It’s important to understand how the factors that influence venture capital investment change throughout a startup’s development. There is a clear gap in the academic literature in learning how the factors that influence startup investment change throughout a startup’s development. Previous approaches to learning the factors that influence venture capital investment in startups have three common weaknesses:

1. Prior work are largely restricted in sample size \cite{croce2016, conti2011, dixon2014, gimmon2010, hoenig2014}. Abundant empirical evidence has suggested that the size of training data eventually becomes more critical than the sophistication of algorithms themselves or even careful feature selection \cite{caruana2008}. Large open databases (e.g. CrunchBase, AngelList) and social networks (Twitter, LinkedIn) offer larger samples than those generally studied in previous works. We expect that using data collected from these sources will lead to the discovery of additional features and higher accuracy in startup investment prediction.

2. Prior work has focused primarily on early-stage investment in startups, primarily equity crowdfunding \cite{beckwith2016, ahlers2015, cheng2016, yuan2016}, angel investing \cite{croce2016} and early-stage venture capital \cite{werth2013}. The functions and objectives of startups change through their development \cite{greve2003}. We expect that the signals that attract investment in these companies will similarly change over time.

3. Prior work has focused on factual company details (e.g. the headquarters’ location, the age of the company, the number of founders) for startup investment predictive models \cite{beckwith2016, dixon2014}, gimmon2010} . Semantic text features (e.g. patents and media) \cite{thorleuchter2012, hoenen2014, yuan2016, wei2009, conti2013} and social network features (e.g. co-investment networks) \cite{wang2016, werth2013, bhat2011, cheng2016, yu2015} may also predict startup investment and performance. We expect that developing a comprehensive model that includes semantic text and social network features alongside factual company features in could lead to better startup investment prediction.

**(0.3) Implications**

This study will develop software that collects and processes information on startups to predict their likelihood of raising investment at different stages in their development. If successful, this study has the potential for scholarly, policy and firm-specific implications. We propose a theoretical framework for startup investment, based on work by Baum & Silverman (2004) \cite{baum2004} and Ahlers and colleagues (2015) \cite{ahlers2015}. Our theoretical framework models startup investment success as a product of two factors: venture quality and investment confidence. We will test this framework with respect to startup development, using cross-sectional and longitudinal analyses. We hope that this study provides interesting insights for entrepreneurs, policy makers, and investors and improve their understanding of the determinants of startup investment, especially for later-stage startups. Ultimately, we hope that this study encourages greater investment in startups.

**(0.4) Signposting**

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

**1. Theoretical Background**

**(1.0) Intro**

**1.1 Technology Startups**

**(1.1.0) Intro**

**1.1.1 Resource Discovery**

Startups are often dependent upon external resources in the early stages of their development because they tend to take time to become profitable. Entrepreneurs require valuable resources such as information, advice, finance, skills and labour when launching startups to be able to realise entrepreneurial opportunities \cite{greve2003}.

**1.1.2 Developmental Lifecycle**

**1.2 Startup Investment**

**(1.2.0) Intro**

**1.2.1 Effect on Startup Performance**

**1.2.2 Investment Assessment**

Potential investors try to evaluate the unobservable characteristics of venture quality by interpreting the signals sent by entrepreneurs as well as potentially a company’s attributes (Connelly, Certo, Ireland, & Reutzel, 2011).

Investors use software to assist them in discovering, evaluating and predicting the performance of startups. Most of this software is not disclosed, though some does occasionally trickle to the media. In 2008, the well-funded startup YouNoodle announced that they had developed software that could predict the future valuation of startups based on analysis of their founding teams \cite{arrington2008}. In 2010, the venture capital firm Kleiner Perkins Caulfield Byers announced that they had developed software called Dragnet that digests App Store data, AngelList entries and Twitter mentions (amongst other datapoints), to surface early-stage startups \cite{geron2013}.

Investors primarily use two approaches to evaluate startup potential: extrapolation of current performance metrics and prediction based on underlying determinants of performance. Dragnet directly evaluates current metrics of startup performance (e.g. app downloads, viral momentum etc.) \cite{geron2013} while YouNoodle analyses determinants of startup performance (in this case, the human capital of the founding team) \cite{arrington2008}. Both approaches have strengths and weaknesses. Dragnet's method of evaluating current performance metrics is easier to implement but YouNoodle's method of evaluating determinants has the potential to be more powerful and explanatory.

**1.2.2 Risk Management**

In a similar context, signaling theory (Spence, 1973) has been used to explain which types of information (board characteristics, top management team characteristics, gender, the presence of venture capitalists or angel investors, founder involvement, etc.) lead investors to invest in start-ups (Ahlstrom & Bruton, 2006; Coleman & Robb, 2014; Cosh, Cumming, & Hughes, 2009; Jääskeläinen, Maula, & Seppä, 2006; Robb & Robinson, 2014). This stream of literature has focused predominantly on the signaling of young start-ups toward angel investors or venture capitalists (Mäkelä & Maula, 2006; Schwienbacher, 2007).

**1.3 Rationale**

**2. Feature Selection**

**(2.0 Intro)**

**<Figure 1 about here>**

**2.1 Venture Quality**

**(2.1.0) Intro**

**2.1.1 Human Capital**

**(2.1.1.0) Intro**

**(2.1.1.1) Company Size**

**(2.1.1.2) TMT Capability**

**2.1.2 Social Capital**

**2.1.3 Intellectual Capital**

**2.2 Investor Confidence**

**(2.2.0) Intro**

**2.2.1 Past Investments**

**2.2.2 Past Performance**

Regarding the selection of the most promising firms, the elevated degrees of uncertainty in the earliest days of a firm's existence may have a negative effect on the VC selection process. Nascent and new firms have a higher failure rate than their more established counterparts (Stinchcombe, 1965; Thornhill and Amit, 2003). Because uncertainty is very high, the value of the firm's future cash flows is difficult to predict even for specialized VC investors, which might limit the positive selection effects in the earliest stages. Moreover, VC investors cannot rely on an operating history of these firms, which might make it impossible to ascertain whether the TMTs function well, if the products under development are competitive, or if customers will purchase and repurchase from the new firm. Thus, the positive selection effects could be lower in the very early stages of the firm's life cycle as compared to later stages.

**2.2.3 Comparable Performance**

**2.3 Feature Summary**

**<Table 1 about here>**

**3. Data Sources**

**(3.0) Intro**

Predicting startup investment is a complex task that c

Appropriate selection of data sources is critical in startup investment analyses because different data sources provide insights into different actors, relationships and attributes \cite{formsma2012, song2012}.

**3.1 Task Properties**

**(3.1.0) Intro**

**<Table 2 about here>**

**3.1.1 Problem**

**3.1.2 Desired Solution**

**3.2 Source Properties**

**(3.2.0) Intro**

**<Table 3 about here>**

**3.2.1 Surveys and Interviews**

Online data collection can be far more efficient than historical methods of collecting data for private company investment analyses because online data sources often have interfaces that allow them to be digested automatically and at scale. Data collected from these online data sources can be publically-available or private and require authentication. Private data is potentially richer and more useful for prediction.

Public pre-formatted data dumps are often available for research purposes, which allows researchers who do not have a computer science background to gain access to these datasets. An alternative to a pre-formatted data dump is a dataset crawler that uses the source's application programming interface (API) to traverse through the dataset, collecting publically-available data. The advantages of such a crawler is that it can selectively collect data from nodes with specific attributes, or collect a random sample (as in \cite{jang2015}), or traverse the data source indefinitely, updating entries as new data becomes available.

**3.2.2 Startup Databases**

There are several startup databases, most are proprietary, CrunchBase and AngelList are large and open….

CrunchBase is an open online database of information about startups, investors, media coverage and trends, focusing on high-tech industry in the United States. It relies on its online community to edit most pages. CrunchBase is a very comprehensive database, with almost complete coverage of startups and investors in the Internet sector, including the relationships between them \cite{alexy2012}. In May 2014, Zhao et al. collected 62,926 investment events between 7,706 VCs and 18,026 startups from 1987 to 2014 on CrunchBase \cite{zhao2015}. However, it has been noted that the CrunchBase corpus is sparse with many missing attributes \cite{xiang2012, zhao2015}.

CrunchBase has two provisions to prevent and remediate inaccurate crowd-sourced entries \cite{crunchbase2014}. First, all users are required to authenticate their CrunchBase accounts with a social media account which allows CrunchBase to verify a user's true identity. Second, every change goes through a machine review, which flags significant or questionable updates for moderation. That aside, further outlier removal will probably be required for analyses using CrunchBase data.

AngelList is a promising new source of startup data, combining the functionality of an equity crowdfunding platform, a social networking site and an online startup database.

As an equity crowdfunding platform, users create profiles for their startups on AngelList, and use the platform to attract investment. Investors use the platform to identify investment opportunities and can invest directly through AngelList, often alongside other investors in investment syndicates.

AngelList is also an online startup database. It has a data-sharing agreement with CrunchBase (which results in significant overlap between the two sources). Importantly, it tracks 'startup roles' (e.g. founders, investors, employees) with a creation time, start time and end time.

Despite being almost entirely absent from the academic literature, AngelList offers substantial data on startups, entrepreneurs and investors which has the potential for some novel insights. For example, Britz et al. \cite{britz2013} analysed the formation of relationships on AngelList (looking at edge-creation time-stamps) to perform a longitudinal study of community growth and development.

**3.2.3 Social Media**

LinkedIn has been used in many online social network studies of entrepreneurship because it is a social network primarily used for professional networking. Unfortunately, as of May 2015, the LinkedIn API no longer allows access to authenticated users' connection data or company data \cite{trachtenberg2015}, making it virtually impossible to use this site for social network analysis, without resorting to semi-automatic HTML parsing techniques (which is against the Terms of Service).

Twitter is a social networking site and micro-blogging site that is often used by entrepreneurs to promote their personal and business brands and rapidly share news and opportunities. Users can send and read public messages (called tweets) of 140 character length. Twitter is a directed network where users are able to follow other users without gaining their permission to do so. Song and Vinig \cite{song2012} found that Twitter network size had a negative relationship with startup survival, perhaps because some entrepreneurs expend too much effort managing this network.

**3.2.4 Patent Filings**

**3.2.5 Financial Reports**

**3.3 Source Evaluation**

Online data sources typically capture a wide spectrum of businesses and entrepreneurs with varying data quality. CrunchBase includes large companies like Google and Ebay as well as small startups founded very recently. It is important to filter entities to some extent before further analyses are performed. One possible strategy is to only analyse companies that have receive venture funding, are in the IT or internet sectors, and are based in the United States, because this is likely to be the largest grouping (as in \cite{alexy2012}).

**<Table 4 about here>**

**4. Learning Algorithms**

**(4.0) Intro**

Machine learning is characterised by algorithms that can improve their ability to reason about a given phenomenon given greater observation and/or interaction with said phenomenon. Mitchell provides a formal definition of machine learning in operational terms: ``A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." \cite{mitchell1997}.

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

The key objective of machine learning algorithm selection is to find algorithms that make assumptions that are consistent with the structure of the problem (e.g. tolerance to missing values, mixed feature types, imbalanced classes) and suit the constraints of the desired solution (e.g. time available, incremental learning, interpretability). In this section, we describe the characteristics of the startup investment prediction task, review common machine learning algorithms, and ultimately determine which algorithms are most likely to suit the characteristics of this task.

**4.1. Task Properties**

**(4.1.0) Intro**

For the startup investment prediction task, we will be focusing on supervised classification techniques. We have a labelled dataset

**< Table 5 about here >**

**4.1.1 Problem**

**4.1.2 Desired Solution**

**4.2 Algorithm Properties**

**(4.2.0) Intro**

The meta-review in Table 6 compares common supervised learning algorithms across characteristics of the two domains mentioned in the previous section: data structure and application problem. The following sections describe each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in relevant applications.

**< Table 6 about here >**

**4.2.1 Naïve Bayes**

Naïve Bayes is a simple generative learning algorithm. It is a form of Bayesian Network that models features by generating a directed acylic graph, with the strong (naïve) assumption that all features are independent. While this assumption is generally not true, it simplifies estimation which means Naïve Bayes is more computationally efficient than other learning algorithms. Naïve Bayes can be a good choice for datasets with high dimensionality and sparsity as it estimates each feature independently. Naïve Bayes has been found to sometimes outperform more complex machine learning algorithms because it is reasonably robust to violations of feature independence, at least with regards to classification (Niculescu-Mizil & Caruana, 2005). However, Naïve Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features (Zhang, 2004). Naïve Bayes was used alongside Logistic Regression, Decision Trees and Support Vector Machines to predict success in equity crowdfunding campaigns on the AngelList data set \cite{beckwith2016}. None of these models performed well. The algorithm that best predicted funded startups was Naïve Bayes with a Precision of .41 and Recall of .19, which means that only 19% of funded startups were classified correctly by the model. The author suggests the poor performance of their algorithms is caused by insufficient features captured in their training set, missing features relating to Intellectual Capital, 3rd Party Validation or Historical Performance. These features are included in the theoretical framework proposed by the current study.

**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 (Hastie et al., 2008), where it is important that specificity is high). It is also simple to update a Logistic Regression model using online gradient descent, when additional training data needs to be quickly incorporated into the model. Logistic Regression tends to underperform against more complex algorithms like Random Forest, Support Vector Machines and Artificial Neural Nets in higher dimensions \cite{caruana2008}. This underperformance is observed when Logistic Regression is applied to startup investment prediction tasks \cite{xiang2012, beckwith2016, bhat2011}. However, weaker predictive performance hasn’t prevented Logistic Regression from being commonly used. Its simplicity and ease-of-use means it is used more casually, often being used without justification or comparative evaluation of its use \cite{gimmon2010, huang2015}.

**4.2.3 K-Nearest Neighbours**

K-Nearest Neighbours is a common lazy learning algorithm. Lazy learning algorithms do not perform explicit generalisation, but compare new instances with instances from training stored in memory. K-Nearest Neighbours is based on the principle that the instances within a dataset will exist near other instances that have similar properties (Cover and Hart, 1967). 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 \cite{kotsiantis2007}. 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 (Ahn & Kim, 2008). K-Nearest Neighbours is attractive in bankruptcy prediction because it can be updated in real-time. By optimising feature weighting and instance selection, the authors managed to improve the K-Nearest Neighbours algorithm to the point where it outperformed the Artificial Neural Network.

**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 (Hunt et al., 1966) and Gini Index (Breiman et al., 1984). 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 \cite{caruana2006}. These shortcomings have been addressed with pruning mechanisms and ensemble methods like Random Forests, respectively. Decision Trees were compared with Naïve Bayes and Support Vector Machines to predict investor-startup funding pairs using CrunchBase social network data \cite{liang2015}. Decision Trees had the highest classification accuracy and the authors suggest they are particularly useful in this application because their reasoning is easily communicated to startups.

**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 \cite{breiman2001}. Prediction is made by aggregating the predictions of the ensemble. The rationale is that while each Decision Tree in a Random Forest may be biased, when aggregated they produce a model that is robust against over-fitting. Random Forests exhibit a performance improvement over a single Decision Tree classifier and are among the most accurate learning algorithms \cite{caruana2006}. 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 \cite{bhat2011}. 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 \cite{breiman2001}.

**4.2.6 Support Vector Machines**

Support Vector Machines are a family of classifiers that seek to produce a hyperplane that gives the largest minimum distance (margin) between classes. The key to the effectiveness of Support Vector Machines are kernel functions. Kernel functions transform the training data to a high-dimensional space to improve its resemblance to a linearly separable set of data. Support Vector Machines are attractive for many reasons. They have typically high accuracy \cite{caruana2006}, theoretical guarantees on limiting overfitting, and with an appropriate kernel they can work well even if data isn’t linearly separable in the base feature space (though this is an issue with a linear kernel). Support Vector Machines are computationally intensive and relatively complicated to tune effectively (compared to Random Forests, for example). Support Vector Machines were compared with back propagated Artificial Neural Networks in predicting the bankruptcy of firms using data provided by Korea Credit Guarantee Fund \cite{shin2005}. Support Vector Machines were found to outperform Artificial Neural Networks at this task, especially because it was on a small data set.

**4.2.7 Artificial Neural Networks**

Artificial Neural Networks are a computational approach based on a network of neural units (neurons) that loosely models the way that the brain solves problems. An Artificial Neural Network is broadly defined by three parameters: the interconnection pattern between the different layers of neurons, the learning process for updating the weights of the interconnections, and the activation function that converts a neuron's weighted input to its output activation. A supervised learning process typically involves gradient descent with back-propagation. Gradient descent is an optimisation algorithm that updates the weights of the interconnections between the neurons with respect to the derivative of the cost function (the weighted difference between the desired output and the current output). Back-propagation is the technique used to determine what the gradient of the cost function is for the given weights, using the chain rule. Artificial Neural networks tend to be highly accurate but are very slow to train and require significantly more training data than other machine learning algorithms. Artificial Neural Networks are also a black box model so it is difficult to reason about their output in a way that can be effectively communicated. Artificial Neural Networks have been rarely applied to startup investment or startup performance prediction tasks, probably because research in this area has used relatively small and low dimensional data sets. As one author put it ``More complex classification algorithms—artificial neural networks, Restricted Bolzmann machines, for instance—could be tried on the data set, but marginal improvements would likely result.” \cite{beckwith2016}. However, the current study seeks to address both of those issues so Artificial Neural Networks may be more relevant.

**4.3 Algorithm evaluation**

We evaluated common supervised learning algorithms for their suitability to the current task, startup investment prediction using the CrunchBase dataset. There is no clear best-in-class algorithm. We expect Random Forests, Support Vector Machines and Artificial Neural Nets to produce the highest classification accuracies. An ensemble of these high-performing methods may also provide an accuracy improvement, though at the cost of computational speed and interpretability. Random Forests could be expected to slightly outperform the other two algorithms due to robustness to missing values and irrelevant features and native handling of discrete and categorical data. However, Random Forests are not highly interpretable so Decision Trees and Logistic Regression might be preferable for early, exploratory analysis of the dataset. Ultimately, however, larger training sets and good feature design tends to outweigh algorithm selection.

**< Table 7 about here >**

**5. Conclusion**

**(5.1) Overview**

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

**(5.2) Rationale**

Venture capital funding for late-stage privately-held startups is approaching all-time highs as investors enter the private markets \cite{nvca2016}. It’s important to understand how the factors that influence venture capital investment change throughout a startup’s development. There is a substantial research gap around accurately predicting startup investment success. Existing approaches in the literature were assessed to have three common limitations: small sample size \cite{croce2016, conti2011, dixon2014, gimmon2010, hoenig2014}, a focus on early-stage investment \cite{beckwith2016, ahlers2015, cheng2016, yuan2016, croce2016, werth2013}, and sparse use of features \cite{ahlers2015, an2015, cheng2016, croce2016, werth2013, gimmon2010, thorleucter2012}. Although individual studies addressed some of these limitations, none attempted to synthesise their findings into a standalone study and piece of software.

**(5.3) Implications**

This study will develop software that collects and processes information on startups to predict their likelihood of raising investment at different stages in their development. If successful, this study has the potential for scholarly, policy and firm-specific implications. We propose a theoretical framework for startup investment, based on work by Baum & Silverman (2004) \cite{baum2004} and Ahlers and colleagues (2015) \cite{ahlers2015}. Our theoretical framework models startup investment success as a product of two factors: venture quality and investment confidence. We will test this framework with respect to startup development, using cross-sectional and longitudinal analyses. We hope that this study provides interesting insights for entrepreneurs, policy makers, and investors and improve their understanding of the determinants of startup investment, especially for later-stage startups. Ultimately, we hope that this study encourages greater investment in startups.

**References**