**Learning the factors that influence investment in startups throughout their development**

**0. Introduction**

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 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 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 (CITATION). 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}.

The trend of startups remaining privately-held for longer has the effect of moving value creation to the private markets. For example, Microsoft’s market capitalisation grew 500-fold following its IPO \cite{microsoft}, but for Facebook to do the same 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 follow value creation to the private markets \cite{nvca2016}. Accordingly, it’s important to understand how the factors that influence venture capital investment change throughout a startup’s development. Previous approaches to learn the factors that influence venture capital investment in startups have three common weaknesses:

1. Prior work is largely qualitative in nature (CITATIONS) or restricted in sample size (CITATIONS). Abundant empirical evidence over the past decade has suggested that the size of training data eventually becomes more critical than the sophistication of algorithms themselves (Banko and Brill 2001; Norvig 2008).

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}. There is evidence that the functions of entrepreneurs and the objectives of startups change dramatically through their development \cite{greve2003} so we expect that the signals that attract investment in these companies will similarly change over time.

3. Prior work has employed numerical operationalizations of financial, managerial, and technological variables in predictive models but largely ignored textual and social network data that is available from media and social media sites. Prior work has demonstrated that social media is effective in detecting events and expressing public opinions (Sakaki, Okazaki, and Matsuo; O’Connor et al. 2010).

There is a clear gap in the academic literature in learning how the factors that influence venture capital investment change throughout a startup’s development. This study could significantly improve our understanding of the determinants of venture capital investment and has the potential to de-risk venture capital and encourage greater investment in startups.

The remainder of the paper proceeds as follows. The next section explores theoretical models of 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 needed 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 summarizes our main results and concludes.

**1. Theoretical Background**

Intro needed

* Underlying factors 🡪 startup performance
* Startup signals 🡪 Underlying factors

**1.1 Startup Performance**

Mobilizing resources to build a new organization is an undertaking laden with uncertainty and unforeseeable hazards. It is also an inherently social process because entrepreneurs must access financial and social capital and other types of resources through relationships with parties beyond the boundaries of their organizations. Because the quality and promise of a new venture is always a matter of some debate, however, the decision of external resource holders to invest time, capital, or other resources in a new organization is one that must be made in the face of considerable uncertainty about the startup’s survival chances and financial prospects.

Many obstacles confront young companies (Stinchcombe, 1965). Startups often lack employee commitment, knowledge of their environment, and working relationships with customers and suppliers. Because they have little operating experience, startups frequently operate using immature and unrefined routines. Startups also tend to be small and so unable to withstand a sustained period of poor performance (Aldrich and Auster, 1986). These perils have led organizational sociologists to conclude that new organizations are highly vulnerable to direct selection, a notion succinctly portrayed as a liability of newness (Hannan and Freeman, 1984; Stinchcombe, 1965). Because startups encounter so many hazards and because they have short-track records by which outsiders can evaluate their potential, there is considerable uncertainty about their value. This uncertainty is compounded for firms established to pursue commercial applications of new technologies (Aldrich and Fiol, 1994). Added to the usual hazards of inexperience, technology startups often require substantial resources to fund early stage and speculative development projects, while revenues cannot be expected until well into the future. New technology is, moreover, by its very nature highly uncertain: undeveloped markets follow unforeseen turns; ‘‘hyped’’ technologies disappear; technologies obsolesce rapidly; and unanticipated ‘‘kinks’’ derail once-promising projects (Tushman and Rosenkopf, 1992). New technology startups are thus particularly risky and uncertain. Given these uncertainties, how do VCs assess the potential of startups and select their investments? VCs spend a great deal of time and effort seeking and assessing signals of a startup’s promise and quality (Amit et al., 1990; Hall and Hofer, 1993). When unambiguous measures of performance do not exist or cannot be observed, investors look for other signs or certifications of future promise and quality (DiMaggio and Powell, 1983; Podolny, 1993). Newly founded ventures must assemble a range of resources and relationships to survive and thrive (Stinchcombe, 1965). Biotechnology startups require access to human, intellectual, alliance, and financial capital (Baum et al., 2000;Walker et al., 1997) and are often portrayed as engaging in a series of ‘‘races’’ to win over desirable managers and researchers, garner valuable patent rights, develop relationships with desirable partners, and obtain the financial resources necessary to support technology development (Amburgey et al., 1996). VC investment is typically viewed as the most critical form of capital (Anderson, 1999; Shepherd et al., 2000); while consistent with the classic signaling literature in economics (Spence, 1974), access to the other forms of capital is seen more as an important signal to VCs of a startup’s future promise (e.g., Stuart et al., 1999). Thus, prior research implicates three broad types of signals that may affect VCs’ assessments of startups: alliance capital, intellectual capital, and human capital.

In evolutionary models of entrepreneurship, entrepreneurs generate variation by founding new firms, pursuing different strategies, and attempting to combine different bundles of assets to do so. Selection is then generated by the decisions of external resource holders to allocate their resources among these firms (Aldrich, 1999). In the entrepreneurial setting, financial intermediaries such as venture capital firms (VCs) have been cited as perhaps the dominant source of selection (Anderson, 1999). VCs affect selection by providing financial resources to cash-hungry startups and by favoring new firms with, or requiring them to adopt, particular strategies, practices, or other characteristics. VCs may also provide management expertise or access to other capabilities that bolster the competitive advantage of startups that they fund (Hellmann and Puri, 2002). Further, because they are perceived to be ‘‘informed agents’’ able to identify particularly promising startups, their provides a certification benefit that can enable the startup to obtain other resources (Megginson and Weiss, 1991). Thus, VCs can affect selection by acting as what we term a ‘‘scout’’ able to identify potential and as a ‘‘coach’’ (Hellmann, 2000) that can help realize it.

Given the importance of financial intermediaries to selection within an organizational population, this is an important gap in our knowledge and our research focus. One way to address this gap is to test the effectiveness of VCs predictive templates by comparing the effects of startups’ characteristics on VCs decisions to finance them with the effects of the same characteristics on future startup performance. If VCs affect selection primarily by picking winners, then the startup characteristics that attract VC investment should also enhance their future performance. If, instead, VCs affect selection primarily by building winners, then the startup characteristics that attract VC investment need not be associated with future startup performance and may even affect it negatively.

**1.2 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).

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).

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}.

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.

**2. Proposed Framework**

We pr Baum & Silverman (2004)

**2.1 Venture Quality**

* Baum & Silverman (2004)
* Podolny (1993)
* DiMaggio & Powell (1983)

**2.1.1 Human Capital**

* Unger, Rauch, Frese and Rosenbusch (2011)
* Doms, Lewis & Robb (2011)
* Robb & Robinson (2014)
* Definition:
  + Identifying and exploiting business opportunities (Shane & Venkataraman, 2000)
  + Definining and realizing a venture’s strategy (Baum, Locke & Smith, 2001)
  + Acquiring additional resources (e.g. financial) (Brush, Greene & Hart, 2001)
  + Building a positive basis for future learning (Ackerman & Humphreys, 1990)
* Signals:
  + Experience and Management skills (Zacharakis & Meyer, 2000)
  + Educational degrees (Levie & Gimmon, 2008) (Backes-Gellner & Werner, 2007)
  + Board composition (Schjoedt, Monsen, Pearson, Barnett, & Chrisman, 2013)

**2.1.2 Social Capital**

* Baum & Silverman (2004)
* Hoang & Antoncic (2003)
* Chung, Singh & Lee (2000)
* Stages of Development
  + Network Founding Hypothesis
  + Network Success Hypothesis (Bruderl & Preisendorfer)

**2.1.3 Intellectual Capital**

* Baumol (2002)
* Cefis & Marsili (2005)
* Baum & Silverman (2004)
* Silverman & Baum (2002)
* Signals:
  + Cohen & Lemley (2001)

**2.2 Investor Confidence**

**2.2.1 Historical Investments**

**2.2.2 Historical 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**

**3. Data Sources**

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 Data Collection**

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. Startup Databases**

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

**3.2.1 CrunchBase**

CrunchBase is an open online database of information about startups, investors, media coverage and trends, focusing on high-tech industry in the United States. It relies on its online community to edit most pages. CrunchBase is a 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.

**3.2.2 AngelList**

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

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

AngelList is also an online startup database. It has a data-sharing agreement with CrunchBase (which results in significant overlap between the two sources). 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.

**3.3 Social Media Platforms**

**3.3.1 LinkedIn**

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).

**3.3.2 Facebook**

Facebook is a friendship-based online social network. Like LinkedIn, it has also been used quite commonly for entrepreneurial online social network studies \cite{song2012, gloor2013}. In contrast with LinkedIn, Facebook profiles involve significant personal / non-business information. Entrepreneurs' networks on Facebook tend to consist of stronger, more personal ties than entrepreneurs' networks on LinkedIn \cite{gloor2013}.

**3.3.3 Instagram**

Instagram is a media-sharing social network particularly popular amongst young people \cite{jang2015}. Accounts are public by default, unless users elect to create a private account. Despite the popularity of Instagram, there is little scholarly work on it, especially in the field of entrepreneurship. However, because of its younger demographic it may provide insights into the extent to which the social behaviour of entrepreneurs changes over generations.

**3.3.4 Twitter**

Twitter is a social networking site and micro-blogging site that is often used by entrepreneurs to promote their personal and business brands and rapidly share news and opportunities. Users can send and read public messages (called tweets) of 140 character length. Twitter is a directed network where users 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.

**3.4 Patent Filings**

**3.5 Financial Reports and Disclosures**

**3.6. Summary**

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}).

**4. Algorithm selection**

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 characteristics**

**< Table 1 about here >**

**4.1.1 Problem characteristics**

**4.1.2 Desired solution characteristics**

**4.2 Algorithm characteristics**

intro

The meta-review in Table 1 compares common supervised learning algorithms across characteristics of the two domains mentioned in the previous section: data structure and application problem. This review is cross-referenced with the relevant characteristics of the current application problem (startup funding prediction) and data set (CrunchBase) to evaluate and rank the candidate learning algorithms. Each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in relevant applications.

**< Table 2 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 using a directed acylic graph, with the strong (naïve) assumption that all features are independent. While this assumption is generally not true, it simplifies estimation. Naive Bayes converges quicker than discriminative models like Logistic Regression, requiring less training data. Naïve Bayes can be a good choice for small datasets with high dimensionality and sparse datasets as it estimates each feature independently. Even for larger datasets, 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 (Beckwith, 2015). None of these models performed well (Mean F1-Score for positive classes = 0.26), especially for startups that ultimately received funding, though Naïve Bayes (AUC = 0.73) and Logistic Regression (AUC = 0.74) produced the best results. The author suggested insufficient features were captured in the training set, particularly features of the company’s product, its marketing content and quality, and the founders’ professional networks. The most commonly selected features in their models were funding history, located in San Francisco, and Twitter presence respectively.

**4.2.2 Logistic Regression**

Logistic Regression is a regression model. Logistic Regression fits probabilities for the response levels using a logistic function. The regression coefficients are usually estimated using maximum likelihood estimation. 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. However, Logistic Regression tends to underperform against more complex algorithms like Random Forest, Support Vector Machines and Artificial Neural Nets in higher dimensions (Caruana et al., 2008). (EVIDENCE – CASE STUDY)

**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 (Kotsiantis, 2007). 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 (Caruana & Niculescu-Mizil, 2006). 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 (Liang & Yuan, 2015). Decision Trees (CART implementation, Breiman et al., 1984) had the highest classification accuracy and the authors suggested they were particularly useful in this application because their reasoning could be easily communicated to startups.

**4.2.5 Random Forests**

Random Forests are an ensemble learning technique that constructs multiple unpruned decision trees from bootstrapped samples of the training data, using random feature selection. Prediction is made by aggregating the predictions of the ensemble. Random Forests exhibit a performance improvement over a single Decision Tree classifier and are among the most accurate learning algorithms (Caruana, 2006). In contrast to a single Decision Tree, Random Forests are less likely to overfit because of bootstrapping at the training data-level and at the feature-level. However, Random Forests are more complex than Decision Trees by nature, taking longer to create predictions and producing less interpretable output. Random Forests were used to predict private company exits using quantitative data from ThomsonOne (Bhat & Zaelit, 2011). 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 (Breiman, 2001). The authors suggested that meta-classifiers such as boosting and bagging also performed well and deserve investigation in future work.

**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 (Caruana & Niculescu-Mizil, 2006), 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 (Shin et al., 2005). Support Vector Machines were found to outperform Artificial Neural Networks at this task, especially because it was on a relatively 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. Neurons are usually classified into input units, which receive information to be processed; output units, where the results of the processing are found; and units in between known as hidden units which allow for high-dimensional approximation. The most common supervised learning algorithm to estimate the values of the neural weights is back-propagation, which efficiently calculates the gradient of a cost function which is then optimised using an algorithm like gradient descent (. 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. (EVIDENCE – CASE STUDY)

**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 3 about here >**

**5. Conclusion**

**References**