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 \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}.

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 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 restricted in sample size \cite{croce2016, conti2011, dixon2014, gimmon2010, hoenig2014}. The size of training data has been shown to become more critical than the sophistication of algorithms themselves or even careful feature selection \cite{caruana2008}. 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. headquarters location, age of company, 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.

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 improves their understanding of the determinants of startup investment, especially for later-stage startups. Ultimately, we hope that this study encourages greater investment in startups.

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