Computational Analytics for Venture Finance

Thomas Rory Stone

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

of the

University College London.

Department of Computer Science

University College London

October 31, 2014

Statement of Originality

I hereby declare that:

- I understand what is meant by plagiarism;
- I understand the implications of plagiarism;
- I have composed this thesis entirely by myself; and
- this thesis describes my own research.

Thomas Rory Stone

University College London

October 31, 2014

Abstract

This thesis investigates the application of computational analytics to the domain of venture finance — the deployment of capital to high-risk ventures in pursuit of maximising financial return. Traditional venture finance is laborious and highly inefficient. Whilst high street banks approve (or reject) personal loans in a matter of minutes It takes an early-stage venture capital (VC) firm months to put a term sheet in front of a fledgling new venture. Whilst these are fundamentally different forms of finance (longer return period, larger investments, different risk profiles) a more data-informed and analytical approach to venture finance is foreseeable.

We have surveyed existing software tools in relation to the venture capital investment process and stage of investment. We find that analytical tools are nascent and use of analytics in industry is limited. To date only a small handful of venture capital firms have publicly declared their use of computational analytical methods in their decision making and investment selection process.

This research has been undertaken with several industry partners including venture capital firms, seed accelerators, universities and other related organisations. Within our research we have developed a prototype software tool **NVANA: New Venture Analytics** — for assessing new ventures and screening prospective deal flow.

We have focused on computational analytics in the context of three sub-components of the NVANA system. Firstly, improving the classification of private companies using supervised and multi-label classification techniques to develop a novel form of industry classification. Secondly, we have investigated the potential to benchmark private company performance based upon a company's "digital footprint". Finally, the novel application of collaborative filtering and content-based recommendation techniques to the domain of venture finance:

- **Multi-label Industry Classification** — we utilise supervised learning techniques (Naive Bayes, c4.5, Random Forests, Support Vector Machines (SVM)) to address the shortcomings of existing schemes (out-of-date, misrepresentation, misinterpretation) and automating the process of classifying private companies.

- **Estimating Private Company Performance** — in collaboration with industry partner Startup Intelligence we analyse the relationship between public indicators of company activities and company financial performance. We define a methodology and evaluation measures using both classification and regression approaches. A methodology for defining private company peer groups is implemented in order to help define relevant peer groups for benchmarking company performance, furthermore, validated through a user study with industry.

- **Top-N Investment Opportunity Recommendation** — working alongside industry partner Correlation Ventures, a US venture capital firm, we apply recommender systems techniques to the venture finance domain. Demonstrating the efficacy of neighbourhood methods (such as item-based k-Nearest Neighbour) and latent factor models (such as Biased Matrix Factorization (BMF) optimised for Bayesian Personalized Ranking (BPR)) in relation to this novel application domain. Our methodology takes advantage of access to historical financing data, provided by Dow Jones VentureSource, a data provider for the venture capital industry, in the task of recommending Top-N investment opportunities.

We conclude by discussing the future potential for computational analytics to increase efficiency and performance within the venture finance domain. We believe there is clear scope for assisting the venture capital investment process. However, we have identified limitations and challenges in terms of access to data, stage of investment and adoption by industry.

Acknowledgements

This research would not have been possible without the feedback and support of many people. Firstly, my academic supervisors Prof. Philip Treleaven and Dr. Dave Chapman. Industry partners and collaborators including Anu Pathria and the team at Correlation Ventures, Thomas Gatten at Startup Intelligence and others including UCL Advances, NACUE, Seedcamp, Silicon Valley Bank, NESTA and The Telegraph. My co-authors Weinan Zhang and Xiaoxue Zhao and co-founders Simon Chan, Donald Szeto and Kenneth Chan. Rob Fitzpatrick for working together on NVANA. Finally, friends and family, especially Christine, Joshua and Sonia.

Contents

List of Figures

List of Tables

Part I

Overview

Chapter 1

Introduction

This introductory chapter covers research context, focusing on early-stage investment and recent developments; the inefficiencies of venture finance; research goals and objectives; and concludes with the overall structure of this report.

The core of this research focuses on software tools (hereafter, referred to as "tools") and computational analytics ("analytics") in relation to the venture finance domain. We have developed a prototype tool, NVANA: New Venture Analytics, and underlying analytics relating to how private companies are classified (Multi-label Industry Classification, see Chapter 5), evaluated (Estimating Private Company Performance, see Chapter 6) and recommended to relevant investors (Top-N Investment Opportunity Recommendation, see Chapter 7), as shown in Figure 1.1.

Figure 1.1: Overview of NVANA

The system is designed to receive input data on companies and investors and output relevant investment opportunities based upon peer-group performance and relevancy for a particular investor. We have also surveyed existing tools (see Chapter 3) supporting the venture capital investment process. However, before we discuss the system design and the underlying analytics we will outline the context and

motivations for such a system and undertaking this research.

1.1 Research Context

Early-stage investment is a key driving force of technological innovation and is vitally important to the wider economy, especially in high-growth and technology intensive industries (such as Life Sciences and Information Technology). Venture finance refers to the financing of private companies through the use of venture capital. Venture capital (VC) is a form of private equity, a medium to long-term form of finance provided in return for an equity stake in potentially high growth companies [BVC11]. Reported global venture capital investments in 2012 totalled US$41.5 billion [Ern13].

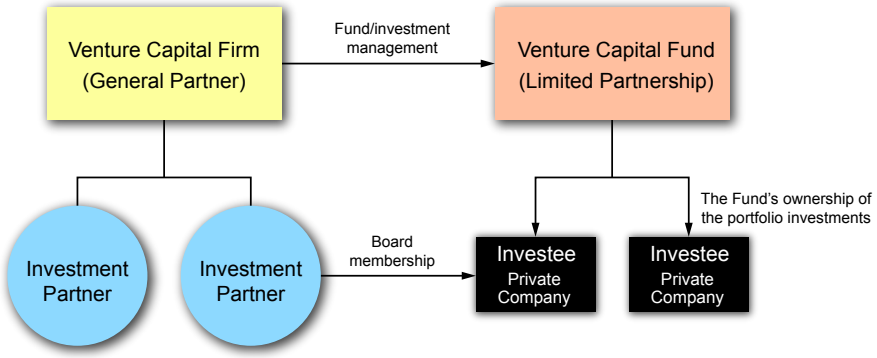


Figure 1.2: Venture Capital (VC) Fund Structure.

Figure 1.2 illustrates the typical structure of a Venture Capital Fund. With some variations a typical fund is managed by a Venture Capital Firm (legally referred to as a *General Partner*) consisting of several investment partners. The fund itself (the *Limited Partnership*) is essentially an investment fund raised from various institutional investors such as pension funds, university endowments and family offices (legally referred to as *Limited Partners*, not shown in our figure or in the scope of our research). Beyond fundraising the main responsibilities of investment partners (also referred to as *General Partners*) are sourcing investment opportunities, making investment decisions and taking board membership to assist the management of investee private companies (also referred to as portfolio companies).

Traditional venture finance is a very labour intensive and time consuming process [FH94]. It involves deploying large amounts of capital and extended due diligence on behalf of the investor. The VC investment process involves several main stages: deal origination, screening, evaluation, structuring (e.g., valuation, term sheets), and post investment activities (e.g., recruiting, financing). Traditionally, investment opportunities are either referred or identified through technology scans [TB84], however, modern information retrieval techniques such as recommender systems have emerged in the past several years as an effective way to help people cope with the problem of information overload [RGFST11, Mon03].

Early stage venture finance is one of the few areas of broader finance yet to be significantly impacted by computational analytics in some form or another. In recent years the traditional venture financing

landscape has shown signs of evolving. Some commentators [Ant12] depict an industry "trifurcating" with i) top-tier firms (e.g., Sequoia Capital), ii) incubators and accelerators (e.g., Y Combinator), iii) and finally, those firms taking a more quantitative approach to funding (e.g., Correlation Ventures). There is potentially a fourth factor in the emergence of entirely new funding sources such as "crowdfunding" which generally operate through online platforms (e.g., AngelList, Crowdcube). Shifts towards more quantitative approaches along with new opportunities for online private investment provide additional impetus and scope for applying computational analytics to this domain. This new domain is quite distinct from existing applications of computational analytics (e.g., Banking, Retail) and particularly recommender systems (e.g., Entertainment, E-commerce) [BFG11] thus representing unique challenges. Whilst there have been some applications of recommender systems to the broader domain of finance, including micro-finance [BS11], there has seemingly been no previous academic research in applying such techniques directly to venture finance.

In this research we seek to develop our application of recommender systems, which is particularly relevant to the screening stage. Although rarely reported a small number of studies show VC investment ratios vary between 1.46 % [BW97] and 3.4 % [Ban91] of investment proposals considered, implying a high rejection rate. This is equally high even for "accelerator" programmes that self-report their own acceptance rates at less than 2% of applications [See]. Our intention is to apply recommender systems with the goal of recommending relevant investment opportunities to VC firms and their investments partners. Importantly we are interested in *early-stage* venture capital investment where detailed financial performance data is rarely available and there is a greater degree of uncertainty compared to later stage private equity deals (such as Mergers, Acquisitions and Buy-outs).

Beyond screening prospective investment opportunities and assessing their "fit" for a particular investor several other tasks in venture finance are reliant on some form of company classification such as identifying peers for competitor analysis or comparables for valuation purposes. With advances in information retrieval, particularly text mining and related techniques, it is possible to envision an improved form of industry classification that more accurately describes the activities and relationships of private companies. We are interested in resolving the shortcomings of existing classification schemes (out-of-date, misrepresentation, misinterpretation) and automating the process of classifying private companies. Furthermore, an alternative representation of private companies activities offers the potential for improved utility (e.g., identifying similar companies, matching investors and relevant investment opportunities) through applying techniques such as recommender systems.

1.2 Problem Definition and Research Justification

In this context we define the following research problems:

- **Identifying relevant investment opportunities.**

As the cost of starting business dramatically decreases investors are faced with and increasingly large number potential businesses and investment opportunities to assess and evaluate. Such a proliferation has led to an "information overload" problem in venture finance. It is important to efficiently identify opportunities which match both an investors' investment thesis, requiring improved classification schemes, and investment criteria, requiring measure of company performance and improved matching or recommendation between investors and investment opportunities in private companies.

- **Existing industry classification schemes are not fit for use.**

Traditional industry classification schemes are not appropriate for VC screening and have the following issues: they are often out-of-date and revised intermittently; companies self-report leading to scope for misinterpretation and misrepresentation; unidimensional schemes lead to ambiguity; and full assignment (i.e., exclusive class membership) doesn't reflect the complexities of the real world. We are interested in resolving the shortcomings of existing classification schemes and automating the process of classifying private companies based upon textual descriptions to improve their efficacy in relation to the venture finance domain.

- **Traditional screening methods are sub-optimal.**

Early-stage investment is characterised by: uncertainty of outcomes for early-stage companies; a lack of reliable data on private company performance; and associated cost of undertaking due diligence. Faced with a large number of prospective investment opportunities (i.e., private companies) venture capital firms and their investment partners require some form of screening. Whilst, referral from trusted sources (e.g., entrepreneurs, accountants, lawyers, other investors) is often used to screen the seemingly infinite number of opportunities seeking further evaluation, relying purely on referral from first or second degree connections is sub-optimal. As an alternative, we believe there is scope for applying techniques such as recommender systems to this domain.

1.3 Research Goals and Objectives

Our main research goals and hypotheses are defined below:

- **Implement a prototype web-based tool, NVANA (New Venture Analytics)** — in collaboration with various partners we designed, tested and implemented a system for assessing new ventures. Both the limitations of our prototype and the potential to extend such a system based upon our work on classification, estimating performance and applying recommender system techniques to this domain are discussed.

- **Apply computational analytics to the novel domain of venture finance** — demonstrating the efficacy of computational analytics in relation to this novel application domain. Our methodology

takes advantage of our access to venture financing data and industry partnerships. We focus on three main areas: *classification*, *performance* and *recommendation*.

- *Multi-label Industry Classification* — Multi-label industry classification will allow for a more useful form of company classification. Traditional industry classification schemes attempt to put companies into discrete classes (or "buckets") unfortunately this isn't representative of the real world. Multi-label industry classification offers a richer representation of private companies activities beyond traditional classification schemes.
- *Estimating Private Company Performance* — Defining private company similarity measures can assist peer identification in order to analyse competitors and compare potential investment opportunities. Utilising our improved form of industry classification we develop measures of private company similarity allowing for benchmarking against a relevant peer-group. Furthermore, we estimate the relationship between potential indicators and actual private company financial performance.
- *Top-N Investment Opportunity Recommendation* — Recommendation systems techniques can complement traditional screening methods. We focus on improving the performance of recommendation models in the task of Top-N investment opportunity recommendation. Additional use cases included finding co-investment partners and identifying the most suitable investors for a company based upon past investment history.

These interrelated applications of analytics to the domain of venture finance align with the three components envisioned for the prototype NVANA system. *Classification* is important for ensuring companies match an investor's investment thesis, *Performance* allows investors to evaluate relative performance of a company against their investment criteria, and *Recommendation* is important for matching relevant companies to investors.

1.4 Report Structure

The report is divided into four main sections *Overview* (see Chapters 1 & 2), *Tools* (see Chapters 3 & 4), *Analytics* (see Chapters 5, 6 & 7) and *Conclusions* (see Chapter 8), with the following chapters:

Chapter 1: Introduction — research context, problem, objectives and structure.

Chapter 2: Venture Finance — recent developments in the domain of venture finance and the scope for applying computational analytics.

Chapter 3: Existing Tools — survey and critical assessment of software tools currently used in industry to support the VC investment process.

Chapter 4: Experimental Tools — description of the design and implementation of a prototype web-based tool NVANA: New Venture Analytics for assessing new ventures.

Chapter 5: Multi-label Industry Classification — design and implementation of the methodology for generating novel industry classification schemes.

Chapter 6: Estimating Private Company Performance — overview of different company similarities measures and estimation of private company performance with industry partner.

Chapter 7: Top-N Investment Opportunity Recommendation — application of recommender system techniques to the venture finance domain including analysis and interpretation of results against intended use cases.

Chapter 8: Conclusions — summary of conclusions, critical assessment, discussion of contribution of research and further work.

Finally, the Appendices provide information to support the main body of the thesis.

Chapter 2

Venture Finance

This chapter covers recent developments in the domain of venture finance and sets out the scope for applying computational analytics.

Early-stage investment is typified by venture capital firms (VCs) who deploy capital towards high-risk ventures. Venture capital has five main characteristics [Met07]:

- is a financial intermediary
- invests only in private companies
- takes an active role in monitoring and helping portfolio companies
- primary goal is to maximise financial return by exiting investments through sale or an initial public offering (IPO)
- invests to fund the internal growth of companies

2.1 Inefficiencies of Venture Capital

Importantly, venture capital is commonly defined as a form of "risk capital" [Bar94] and investment decisions are made under conditions of high uncertainty. Arguably the future outcomes of a venture capital fund's portfolio companies, and by extension the expectation of return, is uncertain in the true Knightian sense [Kni21], meaning it is both unpredictable and unmeasurable. In order to make a financial return a VC fund is reliant on a small number of successful exit events (i.e., acquisitions, initial public offering (IPO)) colloquially referred to as "wins" or "hits". A commonly depicted, if not overly simplified, scenario is that given a portfolio of ten startup companies:

- 3 startups will fail, have gone bankrupt or close
- 3 startups will remain active but will not be very profitable, returning less than the invested capital

- 3 startups will be active and profitable, returning just the invested capital
- 1 highly successful startup will pay the investor a multiple return on all of his 10 investments; through an IPO or acquisition

However, as noted in other studies [SM02] estimates range widely from overly optimistic to the widely acknowledged "one in ten" success rate espoused by industry bodies such as the National Venture Capitalists Association (NVCA). The assertion that "9 out of 10 startups fail" is an oft-quoted yet unattributed statistic. Clearly the true startup failure rate is highly dependent on defining both what constitutes a "startup" and also what we mean by "failure". Prior studies into survival rates in United States using longitudinal data from the U.S. Small Business Administration (SBA) show rates of around 40% survival after 6 years [PK89]. However, this varies widely by industry, for example, survival rates are higher in industries comprised of small innovative firms compared to industries with economies of scale and high capital intensity, whereby failure rates are much higher [Aud95].

Alongside the inherent uncertainty and high risk nature of early-stage investment the historic performance of such investments should give cause for concern. In fact, the upper bound for net risk-adjusted return to the VC industry is zero [Met07]. Therefore investment in VC as an asset class only makes sense if an institutional investor (e.g., pension funds, endowment funds) can consistently select VC funds which outperform the industry average. A recent Kauffman Foundation report [BEP09], based on twenty years as an institutional investor, gave a damning indictment of VC performance:

> "During the twelve-year period from 1997 to 2009, there have been only five vintage years in which median VC funds generated IRRs that returned investor capital, let alone doubled it. It's notable that these poor returns have persisted through several market cycles: the Internet boom and bust, the recovery, and the financial crisis. In eight of the past twelve vintage years, the typical VC fund generated a negative IRR, and for the other four years, barely eked out a positive return."

In fact, several commentators suggest that there is really no such thing as a venture capital "industry" [And11a]. Instead a small handful of funds that perform over time and a much larger mass of disparate financial organisations that under perform.

Despite arguments to the contrary there is still the perception, at least in the United Kingdom, of an "equity gap" [Rig11], as investors have moved to invest in larger, later stage businesses where the risks and uncertainties are less extreme [NMC+09]. This persistent lack of risk capital [GMPR07] for early stage companies is partly due to the high cost of performing due diligence and difficulty in assessing potential risks and returns. Traditional early-stage investment is a labour intensive and time-consuming process [FH94]. There is seemingly a market-failure in the efficient provision of early-stage capital

to prospective new ventures seeking funding [NMC+09]. Investors have limited time and resources to review investment propositions and entrepreneurs often are not "investment ready". There is also a search problem. How do entrepreneurs find the most suitable investors and investors find the most promising opportunities? Given the opacity of information and heterogeneity of participating actors, early-stage investment seems to be a particularly good example of an imperfect market.

Alongside the challenges previously mentioned, in recent years the venture funding landscape has been transformed by the lowering cost of starting a business. The fundamental nature of early-stage, high-technology entrepreneurship has changed since the late 1990s. This change has been promulgated by the widespread adoption of the Internet and inexpensive telecommunications and computing, creating low-cost, global-scale market channels for entrepreneurs. As noted by [Gra13] there is a growing disconnect between venture capital funds, whose traditional model requires them to invest large amounts, and startups that need less capital than they used to. This point is often countered by stating that although is takes less capital to start a company, it still takes a large amount to scale and build a large company. Whether true or not, this still points to a change in the funding requirements of startup companies particularly at the earliest stages of investment.

In reaction to the dramatic changes in relation to starting and growing a new venture the financing environment has also shown signs of evolving. [Ant12] depicts an industry "trifurcating" with i) top-tier firms (such as Sequoia Capital), ii) incubators and accelerators (such a Y Combinator), iii) and finally, those firms taking a more quantitative approach to funding (such as Google Ventures). There is potentially a fourth additional factor in the emergence of entirely new funding sources such as "crowdfunding" [BL10, LS10] further impacting upon the already crowded funding landscape.

2.2 Venture Capital Investment Process

In order to understand how computational analytics may be applied to venture finance we must understand the investment and decision-making process involved. Previous studies in venture finance [TB84] have developed a model of the venture capital investment process, involving deal origination, screening, evaluation, structuring, and post investment activities as the main stages of venture capitalists' decision process.

- **Deal Origination** — the processes by which deals enter into consideration as investment prospects.
- **Screening** — involves a delineation of key policy variables which delimit investment prospects to a manageable few for in-depth evaluation.
- **Evaluation** — involves the assessment of perceived risk and expected return on the basis of a weighting of several characteristics of the prospective venture and the decision whether or not to

invest as determined by the relative levels of perceived risk and expected return.

- **Structuring** — covers the negotiation of the price of the deal, namely the equity relinquished to the investor, plus the covenants which limit the risk of the investor.
- **Post-investment Activities** — covers assistance to the venture post-investment, for example, in the areas of recruiting key executives, strategic planning, locating expansion financing, and orchestrating a merger, acquisition or public offering.

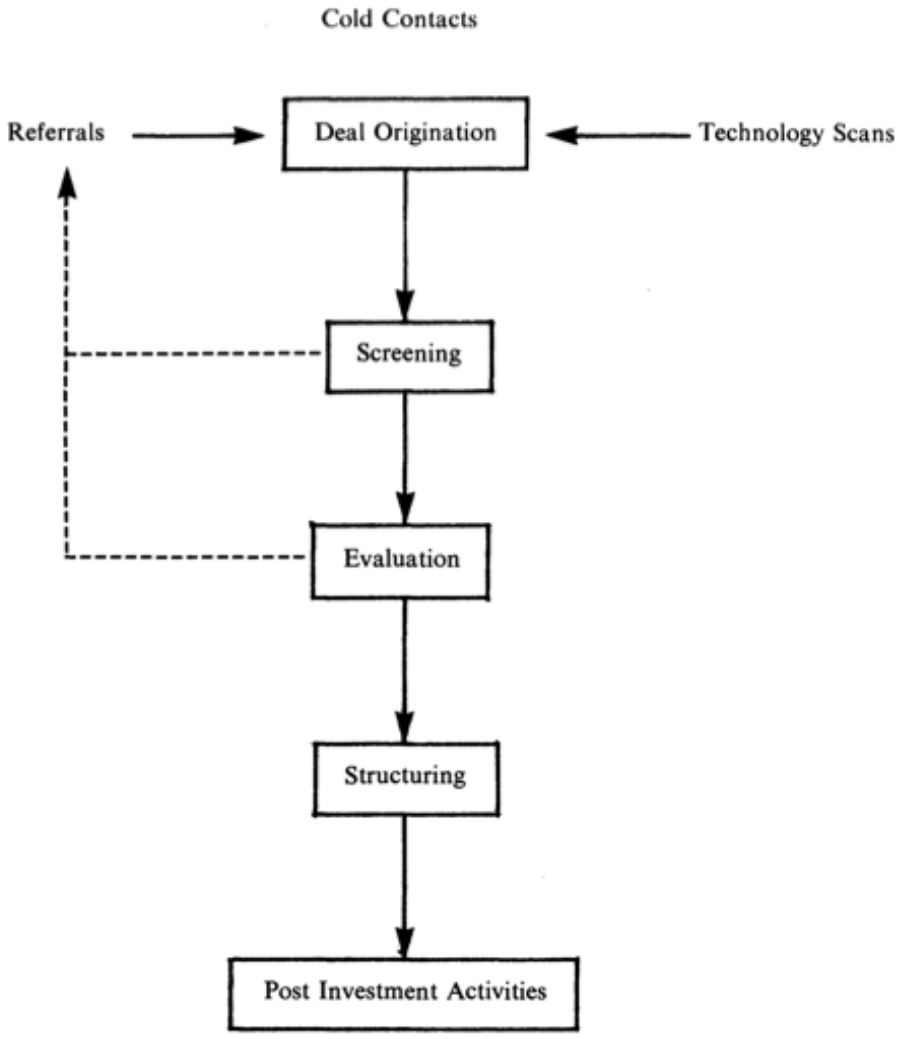


Figure 2.1: Decision Process Model of Venture Capitalist Investment Activity [TB84].

These stages coincide with other research into the general stages of investments [FH94].

2.3 The Quantitative VC

To date only a small handful of venture capital firms have publicly stated their use of analytical methods in their decision making and investment selection process. Three notable examples are Correlation Ventures, General Catalyst Partners and Google Ventures. Correlation Ventures [Cor13] is a US venture

capital firm, which applies predictive modelling to its investment selection process. From their website [1]:

> *"Correlation Ventures has devoted years to building and analyzing what we believe is the world's largest, most comprehensive database of U.S. venture capital financings. Our database covers the vast majority of venture financings that took place over the past two decades, tracking everything from key financing terms, investors, boards of directors, management backgrounds, industry sector dynamics and outcomes. Our selection model informs every investment we make. The data we need is extracted from five readily-available documents provided to us by company management. This data is then supplemented with information from our own knowledge base, so we can rapidly and objectively evaluate any new co-investment opportunity"*

It is important to note that Correlation Ventures only consider co-investment and seemingly the current investors are used as a metric for evaluation. A partner at General Catalyst Partners, devised a method called InvestorRank to assess VC firm connectedness [Sch11]:

> *"Just as Google's PageRank orders search results based on how many links each page gets from other sites, InvestorRank looks at the connections between VC firms. Whenever two VC firms co-invest in the same deal, that creates a bond between them. If one VC firm follows another one in a later round, that boosts the rank of the earlier investor. The more that a VC firm invests in syndicates with other highly ranked firms or even before they do, the higher its InvestorRank. There is some research which suggests that mapping out the network of investors is a better way to predict performance. InvestorRank is not based on previous returns. Rather, it is based on how connected and trusted a VC firm is."*

Google's corporate venture capital (CVC) arm Google Ventures has been noted for running prospective investments through its own algorithms based on historical investment data [Cai11]:

> *"To make its picks, the company has built computer algorithms using data from past venture investments and academic literature. For example, for individual companies, Google enters data about how long the founders worked on start-ups before raising money and whether the founders successfully started companies in the past. It runs similar information about potential investments through the algorithms to get a red, yellow or green light. Google says the algorithms have taught it valuable lessons, from obvious ones (entrepreneurs who have started successful companies are more likely to do it again) to less obvious ones (start-ups located far from the venture capitalist's office are more likely to be successful, probably because the firm has to go out of its way to finance the start-up.)"*

[1] Correlation Ventures - Our Approach — `http://correlationvc.com/approach`

Whilst other venture capital firms have alluded to applying quantitative methods in their appraisal of prospective investment (e.g., e.ventures[2] [Law13], Kliener Perkins Caufield & Byers (KCPB) [Ger13] and Palo Alto Venture Science [Win13]) there is very limited public information about such in-house developments. It is possible that these few investment firms are an early indicator of a shift towards a more quantitative and data-driven approach to early-stage investment. Over a short period of time we have observed dramatic changes in the venture finance domain. In particular the reduced barriers and costs in starting a company and the subsequent increase in potential investment opportunities. This has led to a latent need for improved tools and analytics in order to improve early-stage investment decision making.

2.4 Computational Analytics

The core of this research investigates the application of computational analytics to support the financing of new companies. Our intention is to understand how computational analytics can lead to more efficient and effective financing decisions. To date a limited number of analytical methods have been developed and deployed within industry to provide increasingly relevant insights and support the decision making of investors. These include: modelling expected returns; estimating company valuations; benchmarking peer-group performance; and optimising capital structures.

Computational science and analytics involves using mathematical models and quantitative analysis techniques to study scientific problems through the collection and analysis of increasingly large data sets, and the construction and testing of computer-based models of a system or phenomenon under investigation. In practical use, it is the application of data mining and computer simulation modelling to various scientific disciplines, such as computational finance, computational biology, computational neuroscience, computational chemistry and computational physics amongst other fields. Broadly computational science has two distinct branches:

- **Data Mining** – knowledge discovery that extracts hidden patterns from huge quantities of data, using sophisticated differential equations, heuristics, statistical discriminators (e.g. hidden Markov models), and machine learning techniques (e.g. neural networks, genetic algorithms, support vector machines).

- **Computer Modelling** – simulation-based analysis that tests hypotheses. Simulation is used to attempt to predict the dynamics of systems so that the validity of the underlying assumption can be tested.

Computational statistics and machine learning are used extensively in computational finance, and increasingly in economics. A comprehensive list of techniques spans: symbolic and algebraic comput-

[2] The Daily Gieselmann — http://dailygieselmann.com/

Table 2.1: Computational Finance Taxonomy [YTN10].

Analytical method	Programming technique	Finance applications
Classification	*Rule-based methods*: decision tree learning, first-order learning *Geometric methods*: neural networks, support vector machine *Probabilistic methods*: naive Bayes classifiers, maximum entropy classifiers *Prototype-based methods*: nearest-neighbours classification	Stock selection Bankruptcy prediction Bond rating Fraud detection
Optimisation	Simulated annealing, genetic algorithms *Dynamic optimisation*: dynamic programming, reinforcement learning *Static optimisation*: simplex methods, interior-point methods	Portfolio selection Risk management Asset liability management
Regression	*Dictionary representation*: linear regression, polynomial estimates, wavelet regression *Kernel representation*: k-nearest neighbours, support vector machines	Financial forecasting Option pricing Stock prediction
Simulation	*Stochastic simulation*: Markov chain Monte Carlo simulations *Agent-based simulation*: genetic algorithms, genetic programming	Option pricing Market microstructure

ing, numerical analysis, computational geometry, visualisation and graphics, computational statistics and machine learning.

Machine Learning (or sub-symbolic approaches) refer to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve. Machine Learning further subdivides into *Supervised Learning* and *Unsupervised Learning*. Supervised Learning covers techniques used to learn the relationship between independent attributes and a designated dependent attribute (the label). Most induction algorithms fall into the supervised category (e.g., Decision Trees, Discriminant Function Analysis). Unsupervised Learning covers learning techniques that group instances without a pre-defined dependent attribute. Clustering algorithms are usually unsupervised (e.g., Neural Networks, Self-Organizing Maps (SOM), and Principal Components Analysis).

Focusing on computational finance, Table 2.1 (adapted from [YTN10]) defines various analytical methods, techniques and applications to the broader domain of finance, such as stock or portfolio selection. Whilst computational analytics have been heavily deployed in high finance we believe there is scope for applying related techniques to improve upon investment decision making in the domain of venture finance. We are particularly interested in improving the classification of private companies using supervised learning techniques (see Chapter 5) and both the screening and matching required by venture capital firms using recommender systems (see Chapter 6 and Chapter 7).

Specifically, we rely heavily on data mining, the application of machine learning techniques for both our novel industry classification and the application of recommender systems. In both cases, supervised learning is our main approach, utilising a variety of rule-based, geometric and probabilistic methods (see Section 5.3.5). Whilst unsupervised learning is discussed and used in the preprocessing steps for our industry classification methodology (see Section 5.3.3). We also utilise some computer modelling or simulation in our discussion of findings particularly related to VC and investment partner investment strategies (see Section 7.5).

Limited Use By Industry

Essentially, computational analytics can be viewed as the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Whilst clearly related, "analytics" goes beyond purely data "analysis" (i.e., descriptive statistics) informing decision making. Therefore, for our purposes, We define "computational analytics" as the use of analytical techniques and methods (e.g., classification, regression, clustering) to inform decision making.

In terms of the applicability of computational analytics to venture finance it is interesting to observe the opinions and reported commentary of prominent venture capitalists. When asked whether data-driven principles can be applied to investing in tech startups some well known and respected investors responded [Fro11].

Fred Wilson of Union Square Ventures said:

> *"We have not been able to quantify it. We haven't even tried. Although I am sure someone could do it and they might be very successful with it. To us, the ideal founding team is one supremely talented product oriented founder and one, two, or three strong developers, and nothing else. The supremely talented product oriented founder should have been obsessed about a product area/idea for a long period of time and just has to build something to satisfy their passion/curiosity. That's about it. Joshua Schachter/Delicious, Jack Dorsey/Twitter, Dennis Crowley/Foursquare are the iconic examples of this kind of person in our portfolio."*

Chris Dixon of Founders Collective and now Andreessen Horowitz said:

> *"One of the main activities of good investors is trying to find "accurate contrarian theses" about what make good startups, markets, founders etc. So there is a lot of Moneyball-esque activity. I've seen a few attempts to do it quantitatively (I recall an academic paper on it and also some studies done internally at VCs) but I think those are often flawed because the quantitatively measurable things are either obvious (e.g. founders who sold their last company for a boatload of money are more likely to be successful than founders who failed), irrelevant, or suffer from overfitting' (finding patterns in the past that don't carry forward in the future). Personally, I think the biggest "Moneyball" opportunities in seed investing are around the processes used. For example, I think the format of spending a few hours getting pitched' is a deeply flawed process for getting to know whether a first time founder will be successful. You can think of Y Combinator as an example of trying a new process. I'm personally constantly experimenting with different getting to know founders' processes."*

Paul Graham of Y Combinator said:

"I know of no reputable investor who invests based on data. I once heard of someone who planned to, but I forget who it was; probably nothing came of it. [...] We are the far opposite end of the spectrum from an analytical approach. We decide based on gut feel after a 10 minute convo. It may seem ironic that we who have the most data make the least use of data. But perhaps not: perhaps it's because we have so much data that we know it all comes down to the personalities of the founders. Or maybe we're just lazy."

Other investors including Roger Ehrenberg of IA Ventures also commented suggesting there it is important to distinguish between *security* and *investment* selection [Ehr11]:

"When I speak of security selection, I'm specifically referring to the choice of founders with whom to work. [...] When I speak of investment selection, I'm referring to the choice of investors with whom to invest [...]

I was on the quantitative end of Wall Street for more than 15 years, and have been doing seed stage venture investing for more than 7 years. Both experiences have shaped my perceptions concerning the application of data and quant tools to choosing the best companies with which to partner [...] My own experience has shown that the impact of people and team is far greater than market and product, the last two of which better subject themselves to quantitative assessment than the former. The degree of variability concerning the range of possible outcomes is much more heavily impacted by how a team interacts, innovates and executes than the perceived market opportunity. Modeling markets is often a worthwhile endeavor; modeling human dynamics in seed stage investing is generally not. The notion of pattern recognition - of accumulation of subjective experience - is in my opinion far more constructive than any heuristic. [...]

Investment selection is another matter entirely. In fact, a wide swath of the angel and Micro VC segment are based upon the efficacy of social proof. This speaks more to the success and credibility of those also investing in a security than a deep understanding of the security itself. My friend Bryan Birsic made the comment that Correlation Ventures (CV) takes quantitative approach to VC. Given the framework above, I'd say that CV take a quantitative approach to investment selection - not security selection. This is a very important distinction. The CV model does not allow them to lead rounds, because they are not claiming that their algorithms work on a stand-alone basis; they require the quantitative version of social proof as key inputs into their model for whether or not to invest in a company [...]

In short, I do think that quantitative methods such as those used by CV are likely to be effective as they are systematizing and institutionalizing the notion of social proof. But it is not a substitute for security selection; that needs to be done by firms relying on keen

assessments of people and their ability to execute a vision, adapt in the face of change and persevere against all odds. At least at this point in time, you can't model that."

It is important to note that whilst each of these investors would be deemed *early stage* their typical investment size is often in millions of dollars (i.e., Series A) usually in companies with revenues or a certain level of traction (i.e., active users, subscribers, paying customers, etc). That is with the exception of Y Combinator which is an example of an "accelerator" programme and literally invests in teams with an idea, and recently even in strong teams without a specific business idea [Y C12], investing around $20,000 usually for 6 or 7% in equity. This brings us on to the importance of stage of investment.

Whilst there may be merit in applying analytical techniques to venture finance, early-stage companies have insufficient track record and performance data is either sparse or non-existent, software tools are nascent and use of analytics in industry is limited. Commonalities can be identified between the early-stage investment and applications for short-term loans becoming automated and highly efficient. It is still true that it may take an early-stage VC months to put a term sheet in front of a fledgling firm yet banks can approve personal loans in a matter of minutes. However it is possible to take the comparison too far when we are dealing to two quite fundamentally different forms of finance. The following are taken for granted and simply non existent in the case of venture finance: a detailed credit history and credit scores (e.g., FICO score) $\longrightarrow$ limited operating history; short return period $\longrightarrow$ long return period; small loans $\longrightarrow$ large equity investments; minimise bad debt $\longrightarrow$ selecting winners. Essentially, early-stage venture finance is characterised by an inherent difficulty in assessing risk and potential return. There is a clear link between stage of investment and quality of data, and, therefore, ability to perform meaningful analysis.

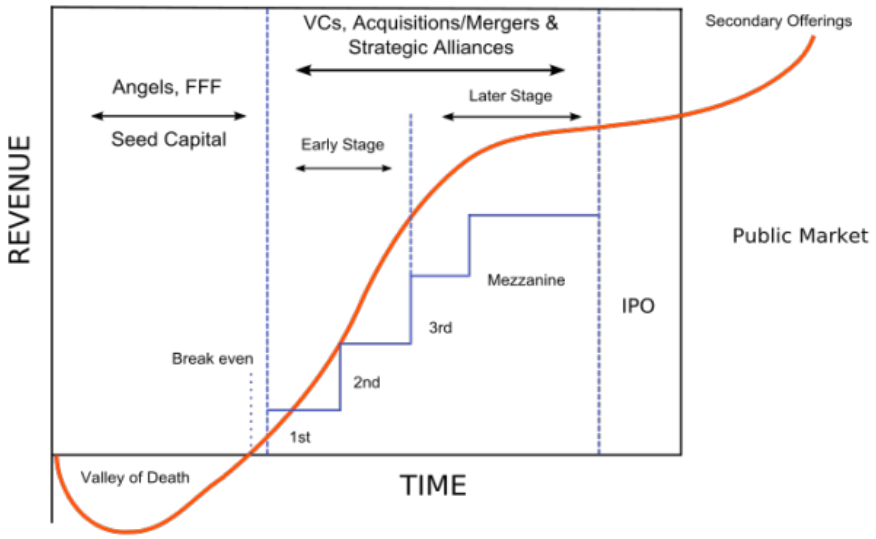


Figure 2.2: Startup Financing Cycle [Sta12].

Figure 2.2 outlines the traditional venture funding cycle and illustrates the different sources and

stages of funding in relation to time and revenue. Different analogies for funding include, "gears" [Gra05] or "ladders" [oT07], to convey the fact that ventures often go through several rounds of funding. Investment of risk capital can be segmented into stages of funding. The British Venture Capital Association (BVCA) commonly defines the stages as early-stage, growth and management buy-out (MBO) or buy-in (MBI) [BVC11]. Similarly, the European Venture Capital Association (EVCA)'s Yearbook [Eur11] classifying funding rounds from seed through to expansion. Although definitions vary, "early-stage" is generally perceived to be less than £2m. As we a mainly concerned with early-stage funding we shall ignore later stage private equity financing, such as mezzanine, bridge or MBO/MBI. Different type of investor will look to invest at different stages of investment; those associated with early-stage funding generally include public or government agencies, universities, seed accelerators, angel investors and some venture capital firms. Clearly, other sources of finance such as organic growth (e.g., sales revenue) and debt financing (e.g., bank loans) have not been mentioned, this is because there are not typically assumed to be forms of "risk capital", however, they often provide necessary financing for many businesses. As discussed earlier, a shortage of early-stage funding has been a particularly acute problem in Europe where investors seemingly focus on later stages of investment (i.e., growth, management buy-out/in) [GMPR07, NMC+09, Rig11]. We believe this persistent lack of risk capital for early stage companies is partly due to the high cost of performing due diligence and difficultly in assessing potential risks and returns. These are essentially problems that could be mitigated through the appropriate use of technology, which have the potential to support such early stage investment.

The following section covers Tools and applications of computational analytics in practice. With a survey of existing tools, the design and implementation prototype tool NVANA, and finally, the discussion of future applications in the domain of venture finance.

Part II

Tools

Chapter 3

Existing Tools

This chapter provides a survey and critical assessment of software tools currently used in the venture capital industry.

We are particularly interested in existing computational analytics and software tools used in the domain of venture finance. A survey of software tools used was undertaken to get a better understanding of the application and use of analytics by the VC industry.

To borrow a quote from a highly respected entrepreneur and now prominent venture capitalist, "software is eating the world" [And11b]. Depicting how entire industries are being reimagined and consumed by software companies, such as with retail (e.g., Amazon), entertainment (e.g., Netflix), music (e.g., Apple), telecoms (e.g., Skype) and advertising (e.g., Google). Perhaps the next industry to be consumed might be venture capital itself. Consequently, the innovation and advancements in the venture finance domain are often manifest in software rather than in traditional academic sources, such as journals and conference proceedings.

The previously outlined investment process model (refer to Figure 2.1) provides a useful framework for observing different activities and therefore potential application of analytics within the context of venture finance. It is also useful for structuring our survey of software tools (Deal Origination, Screening, Evaluation, Structuring, Post-investment Activities) in order to review the different offerings.

3.1 Deal Origination

Deal Origination is the processes by which deals enter into consideration as investment prospects [TB84]. Tools relating to Deal Origination generally can be classified as either: a) *Investment Networks*, directly connecting entrepreneurs and investors; b) *Databases*, providing data and key information about companies; and c) *Market Intelligence* platforms, which allow investors to visualise trends and opportunities within particular markets and industries.

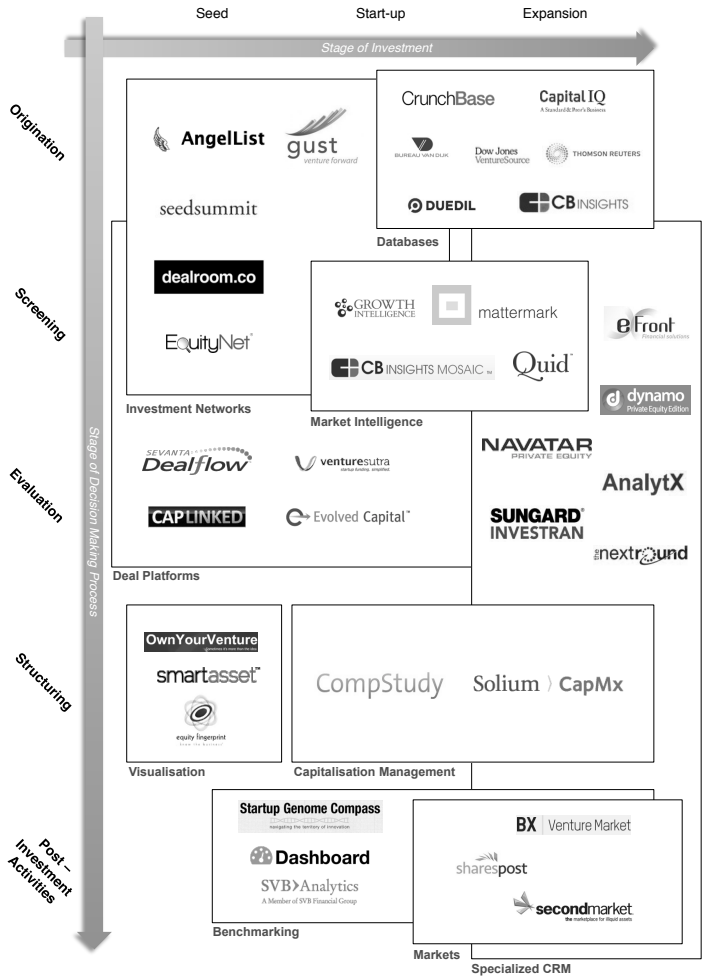


Figure 3.1: Software Tools for VC Investment Process.

3.1.1 Investment Networks

Investment Networks are essentially fully-fledged social networks linking entrepreneurs and investors. Examples include AngelList, Gust, SeedSummit and Dealroom. AngelList[1], founded by those behind the popular Venture Hacks blog and supported by the Kauffman Foundation, allows entrepreneurs to post their funding needs to essentially a social network for start-ups, which in turn is reviewed by investors using the service. AngelList claims to effectively harness "social proof" and allows potential investors to stay abreast of investment opportunities. Gust (formerly Angelsoft) is officially supported by several global angel networks. SeedSummit[2] is a similar concept to AngelList launched by the seed accelerator Seedcamp in order to bring more visibility to European angel investors. These Investment Networks mainly operate in the seed stage of investment and often allow nascent entrepreneurs and start-ups to secure their first outside capital. Other Investment Networks include Dealroom[3] (formerly NOAH Insider), which is also focused on Europe and is linked to a well-known NOAH internet industry conference.

[1] AngelList — `https://angel.co/` [2] SeedSummit — `http://www.seedsummit.org/` [3] Dealroom — `http://dealroom.co/`

3.1.2 Databases

Popular Databases used in industry include Dow Jones' VentureSource[4] with accompanying news source VentureWire; Standard & Poor's Capital IQ[5], combining company information and market research with tools for fundamental analysis, idea generation, and workflow management; Bureau van Dijk has several national and international database offerings such as FAME[6], which covers private companies in United Kingdom and Ireland; and Thomson Reuters' ThomsonONE[7] (formerly VentureXpert and Thomson One Banker), which is endorsed by the National Venture Capital Association (NVCA) as the official United States venture capital activity database. Finally, CrunchBase[8] described as "the free tech company database" which is maintained by the web-based technology publication TechCrunch (owned by AOL).

More recent additions include: CB Insights[9] which offers an alternative to industry data providers allowing investors to source and gather market intelligence on prospective investments, acquisitions, co-investors and financing trends. Building upon their database offering they are also planning to launch CB Insights Mosaic (see Market Intelligence); and finally, Duedil[10] offers a web-based platform for performing due diligence online, enhancing traditional company information with social data, currently available in the United Kingdom and Ireland only.

3.1.3 Market Intelligence

Market Intelligence platforms generally utilise large sets of data in order to give insight and provide information on market dynamics and participants. In relation to the VC industry, offerings include: Growth Intelligence[11] (formerly Startup Intelligence), which as well as producing bespoke reports on start-up hubs, recently launched a business intelligence solution; Quid[12], founded by members of the Younoodle team (see Section 2.4), which attempts to capture, structure, and visualise vast amounts of open information, to help organisations make more informed decisions, harvesting data as diverse as patents applications, NASA grants and FDA approvals; and Mattermark[13] which intends to quantify the growth signals and market data of private companies around the world. They provide a "Mattermark Score" which is an index based on the public"growth signals" they capture (Website traffic, app store rankings, employee count, time since last funding, total funding amount, co-investors, social media). Similar scoring and rankings are offered by Inkwire[14] and Datafox[15] amongst others. Finally, the previously mentioned CB Insights Mosaic software intends to algorithmically assess the health of private companies.

[4] VentureSource — https://www.venturesource.com/ [5] Capital IQ — https://www.capitaliq.com/ [6] FAME — https://fame.bvdinfo.com/ [7] ThomsonONE — https://www.thomsonone.com/ [8] CrunchBase — http://www.crunchbase.com/ [9] CB Insights — http://www.cbinsights.com/ [10] Duedil — https://www.duedil.com/ [11] Growth Intelligence — http://www.growthintel.com/ [12] Quid — https://quid.com/ [13] Mattermark — http://mattermark.com/ [14] Inkwire — http://inkwire.io/ [15] Datafox — http://www.datafox.co/

There are numerous tools that allow investors to develop a pipeline (or "deal flow") of potential start-up companies. Alongside established providers new entrants such as the Investment Networks and Market Intelligence tools, have started to see usage. Despite such advances a majority of qualified deals are still originated through traditional means. Hence traditional "offline" networks and referral via third parties (e.g., existing portfolio companies, accountants, lawyers, etc.) are still key competitive advantages. The tools discussed above often supplement or expedite the process by increasing the available information about potential investment opportunities.

3.2 Screening

Screening involves a delineation of key policy variables which delimit investment prospects to a manageable few for in-depth evaluation [TB84]. Functionality relating to Screening can generally be viewed as a component of tools related to Deal Origination or *Investment Platforms*. Tools relating to Screening sift funding applications, identifying those that warrant further study.

Whilst Screening is not necessarily an activity where tools exist in their own right, it is nevertheless an important step in the investment process. Investors often are faced with hundreds, if not thousands, of prospective investment opportunities and being able to reduce that number to a more manageable subset for detailed evaluation is key. To some extent the previously discussed Databases offer the ability to screen prospective investments based upon geography, industry and other pertinent criteria. Screening components allow entrepreneurs to submit detailed information about their business and then support selection based on a particular investor's investment criteria. This screening often occurs prior to submission, for example, Gust[16] (formerly Angelsoft) offers an "investor search engine", ensuring companies only attempt to seek funding from relevant investors, by matching according to location, industry and other relevant criteria.

3.2.1 Investment Platforms

Investment Platforms, are systems that span multiple phases of the investment process. These subdivide into: a) *Deal Platforms* which help investors to manage their deal flow of prospective investments, and b) *Specialised CRM Systems* which are used by to track deal flow and contacts across large organisations

Deal Platforms

Deal Platforms offer a solution for investors to manage their deal flow of prospective investments. Ultimately, these solutions provide a platform for connecting entrepreneurs and investors through the process of the early stages of due diligence. These include Sevanta Dealflow[17] offering custom hosted workspaces; CapLinked[18] providing cloud-based management of capital raising and asset sales; and

[16] Gust — `https://gust.com/` [17] Sevanta Dealflow — `http://mydealflow.com/` [18] CapLinked — `http://www.caplinked.com/`

EquityNet[19] combining business plan software, tools for analysis and an investment network.

Specialised CRM Systems

Specialised CRM (or Customer Relationship Management) Systems are particularly relevant to later stages of investment, not only in expansion but beyond in the realms of later-stage private equity. In general, such systems are used by to track deal flow and contacts across large organisations. Hence, the relevance to later-stage investors, which are generally larger operations. The systems effectively enable partners and employees to leverage each other's interactions and relationships. Other uses include logging meetings, maintaining client contact details and for other marketing purposes. There are many solutions available to the private equity industry, these are often CRM-type systems and have vast functionality bespoke to the needs of the later-stage private equity investors. Notable offerings include FrontInvest[20] by eFront; Investran[21] by SunGard; plus others such as SS&C (formerly The Next Round) and DealDynamo by Netage Solutions; all of which offer a number of products and services in relation to wider alternative asset management, such as investor relations and fund management, across the full lifecycle of private equity fund activities; finally, Navatar Private Equity Cloud[22] a custom version of the popular Salesforce.com CRM platform.

Screening components provided by CRM (Customer Relationship Management) Systems, are used to track deal flow and contacts across large organisations. These systems commonly offer filters and custom rules in order to organise and streamline dealflow. Investran by SunGard enhances deal management through advanced pipeline management, including the ability to target and source deals within a specific industry or geographical region and organise deal pipeline by both status and stage. Clearly, Screening components are an integral part of the investment process and therefore exists in various different manifestations (e.g., Databases, Investment Platforms). Generally they allow filtering based on common characteristics, such as location, industry and stage of a company, however, many tools allow further customisation and advanced functionality in this respect.

Screening is necessary to reduce the relatively large number of potential deals to a more manageable few, which are evaluated further. Investors commonly define certain investment criteria characterising the type of investment opportunity they will consider. For example: the size of required investment, the industry sector and the geographical location amongst other factors. Traditionally this screening would be a completely manual process, with business plans being reviewed by associates in order to decide whether they meet criteria for a certain fund and merit further consideration. More recently, with the use of web-based applications and software tools, the task has the potential to be partially automated. This can be done explicitly by outlining the criteria as a sort of checklists for those seeking funding. Also,

[19] EquityNet — `https://www.equitynet.com/` [20] FrontInvest — `http://www.efront.com/`
[21] `http://www.sungard.com/` [22] Navatar Private Equity Cloud — `http://www.navatargroup.com/`

implicitly, by filtering out investment proposals already received and under review.

In relation to Screening, Deal Platforms and specific private-equity CRM Systems have gained some acceptance, but it is still very common to find generic software tools including: spreadsheets (e.g., Microsoft Excel); databases; customer relationship management solutions (e.g., Salesforce.com); and custom built file management systems; for the purpose of managing and screening deal flow.

3.3 Evaluation

Evaluation involves the assessment of perceived risk and expected return on the basis of a weighting of several characteristics of the prospective venture and the decision whether or not to invest as determined by the relative levels of perceived risk and expected return [TB84]. The main tools in relation to Evaluation are typically found as components of the Investment Platforms previously discussed. The Evaluation component usually centres on some form of secure workspace or "deal room" in which to store, share and evaluate various important documents (e.g., business plans, financial forecasts, legal documents) related to a particular company or deal. Although typically found as components Investment Platforms.

The Evaluation component offers a solution for investors to manage their due diligence of prospective investments. For example, the previously mentioned Gust, offers a secure deal room, which is automatically created for each deal allowing collaboration privately with all interested investors and authorised third parties. This provides a secure, private platform where investors can collaborate, quickly gauge deal interest and browse ratings and reviews. At least with regards to Gust, the deal room can be described with two main functions: i) A deal-based message board to keep all discussions about the deal organised in one place, ii) A document management system, to allow investors to easily access the deal's business plan, financials and other documents. In a similar vein CapLinked offers a private deal room for tracking communications amongst investors and sharing documents for the purpose undertaking due diligence.

Such tools are commonly used to track deal flow and contacts across larger organisations. They also allow investors to manage all relevant information for prospective investments in a collaborative manner. Through the use of private and secure data rooms, Investran by SunGard allows investors to track financial data (e.g., revenues, debt, pre-and post-money valuations), full capitalisation tables and generate user-defined performance metrics (e.g., multiples). Evaluation components allow investors to capture relevant information regarding prospective investments and perform collaborative evaluation, via deal rooms, thus supporting the due diligence process. Clearly, dependent on the stage of investment (i.e., Seed through to Expansion) the amount of information, particularly in relation to historical performance, financial statements and details of prior investments will be available only at later stages of investment.

3.4 Structuring

Structuring covers the negotiation of the price of the deal, namely the equity relinquished to the investor, plus the covenants which limit the risk of the investor [TB84]. Tools relating to Structuring can be generally classified as either: a) *Visualisation* tools, which provide simple ways to visualise a deal and equity ownership; or b) *Capitalisation Management* tools, which offer richer visualisation and modelling capability in order to evaluate different deal structures and potential outcomes.

Structuring is the final stage before investment and involves pricing of a particular deal, which until recently had little or no specific tools, relying more on spreadsheet modelling. Now Visualisation tools provide a simply way to envision a deal and equity ownership; whereas Capitalisation Management tools, offer richer visualisation and modelling capabilities in order to evaluate different deal structures and potential outcomes.

3.4.1 Visualisation

There are a small number Visualisation tools, which can be used to assist in structuring financing and pricing deals. Equity Fingerprint[23], provides a free service relating to equity-mapping and offers relevant resources for businesses, with the intention of understanding how investment and equity dilution relates to the growth of a venture. Other visualisation tools, OwnYourVenture[24] and Captable.io[25], both offer free web-based equity financing and dilution calculator. These tools are potentially more relevant to entrepreneurs raising funding and serve as decision aides rather than fully functioning software solutions.

3.4.2 Capitalisation Management

Structuring and company valuation, especially in relation to early-stage investment, is often described as much an art as it is a science. In theory, there are numerous approaches to company valuation, including revenue-based (e.g., Discounted Cash Flow); cost-based (e.g., Book Value); market-based (e.g., Comparable Analysis); or hybrid approaches (e.g., First Chicago Method) commonly used in the venture capital industry.

In relation to Capitalisation Management tools, a market leader is Solium CapMx[26] (formerly SVB Analytics CapMx and formerly eProsper) equity compensation management software, which allows companies to manage private company stock options, preferred stock, restricted stock and warrants. Whilst this service is offered by other firms, through the acquisition of CapMx now Solium are able to offer this service through a web-based software solution. A related service is CompStudy[27], a comprehensive report of salary and equity data for senior management at private companies. It includes data on more than 3,000 companies and 25,000 executives compensation.

[23] Equity Fingerprint — `http://www.equityfingerprint.com/` [24] OwnYourVenture — `http://www.ownyourventure.com/` [25] Captable.io — `https://captable.io` [26] Solium CapMx — `http://capmx.solium.com/` [27] CompStudy — `https://www.compstudy.com/`

Until very recently, few standalone tools existed in relation to structuring or pricing a deal, perhaps due to the prevalence of custom spreadsheet modelling (e.g., using Microsoft Excel) and occasionally add-ons allowing more complex simulation modelling (e.g., @RISK by IBM or Crystal Ball by Oracle). Furthermore, this is compounded by the fact that investors may have different approaches to performing valuation. Tracking the complex capital structures of different investments, including complexities such as "preferences", "warrants", "anti-dilutions provisions", "ratchets" and "clawbacks", over several rounds of investment, can quickly become a very unenviable task. More recently tools such as those covered under the category Capitalisation Management have become available to deal with this very problem. Surprisingly, one of the key drivers of adoption for such tools has been regulation (particularly in the US), which forces investors to disclose detailed information on their existing portfolio of investments.

3.5 Post-investment Activities

The term post-investment is used to cover software assistance to the venture in the areas of recruiting key executives, strategic planning, locating expansion financing, and orchestrating a merger, acquisition or public offering [TB84]. Within the scope of this survey, tools relating to Post-Investment activities can be generally can be classified as either: a) *Benchmarking* tools, which allow investor to compare company performance against a relevant peer-group; or b) *Markets* for private companies, which allow trading of illiquid assets within a market of potential investors.

3.5.1 Benchmarking

Benchmarking tools allow companies to benchmark their performance against their competition and peer-group. In relation to seed stage companies with little financial track record the Startup Genome Compass[28] offers a is a simple benchmarking tool for entrepreneurs to evaluate their progress against other start-ups and make more informed product and business decisions. For later stage companies interest in benchmarking financial performance against a relevant peer-group then SVB Analytics Benchmarking[29] offer access to proprietary aggregate performance data for similar companies. Dashboard.io[30] which grew out of the 500 Startups accelerator and community provides investors early-stage data they need to make better portfolio decisions based upon comparing and benchmarking company progress. It also intends to better connects startups with relevant communities, events and tools.

[28] Startup Genome Compass — `https://www.compass.co/` [29] SVB Analytics — `http://www.svb.com/analytics/` [30] Dashboard.io — `https://dashboard.io/`

3.5.2 Markets

Markets tools such as the private markets Sharespost[31] and SecondMarket[32] (formerly Restricted Stock Partners) offer a marketplace to buy and sell shares in private companies. Exitround[33] is the private, anonymous marketplace for buyers and sellers of technology companies from the small to mid-market. More recently NASDAQ launched BX Venture Market[34] as a separate listing venue from main NASDAQ exchange in order to allow smaller companies with lower liquidity to attract investors and raise capital.

There are almost limitless post-investment activities that may be relevant to investors depending on their approach, including recruitment, financing, partnerships, monitoring performance and everything in between. The two categories of tool covered here are both fairly new innovations. Benchmarking tools offer clear value in relation to the ongoing oversight of portfolio companies against peers and competition. Meanwhile, secondary private Markets have proliferated over the past few years (perhaps due to the weak prospects of public offering on established exchanges) and, so far, have not been subject to any official review by regulators.

3.6 Discussion

Whilst the tools covered provide support for venture capital process, there has been *some* adoption (e.g., Specialised CRM Systems by later stage private equity firms) but in the earlier stages of investment there has been relatively low utilisation of any of the solutions outlined. Perhaps due to the overhead of implementing such a system, the perceived benefits may be realised only by larger organisations with larger prospective deal flow.

Whilst it is possible to envision a complete platform attempting to offer functionality across the entire investment process, the development of specialised tools focusing on specific aspects of the investment decision-making process seems more likely. As is already the case in relation to various tools which offer application programming interfaces (APIs), we observe specialisation with greater interoperability between Deal Platforms, Databases, Markets and other such categories of tools. Finally, it is important to note that, despite the large number of providers and tools covered in this survey paper, actual adoption by industry is relatively low, and it is difficult to identify current penetration and significance due to the lack of publicly available information.

The following Chapter introduces a prototype software tool NVANA (New Venture Analytics) for assessing new ventures and screening prospective deal flow. Including the limitations of our prototype and potential extensions.

[31] Sharespost — `http://sharespost.com/` [32] SecondMarket —- `https://www.secondmarket.com/` [33] Exitround — `http://exitround.com/` [34] BX Venture Market — `http://www.bxventure.com/`

Chapter 4

Experimental Tools

This chapter focuses on the development of a prototype software tool to support the venture capital investment process. We discuss the design and implementation of a web-based software tool, NVANA (New Venture Analytics) which was developed to assist in the appraisal of early-stage ventures.

.

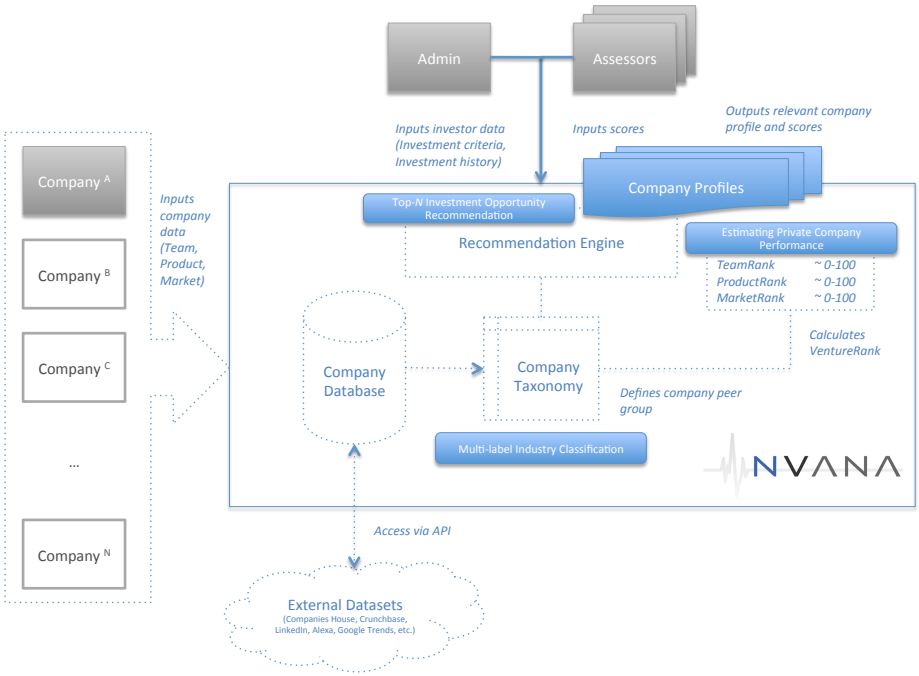


Figure 4.1: Overview of NVANA.

As part of ongoing research a prototype web-based tool for managing applications for funding (e.g., business plan competitions) and assessing new ventures was designed, tested and implemented. Supported by University College London (via UCL Advances), the National Association of College and

University Entrepreneurs (NACUE) and the China Innovation and Development Association in the UK (CIDA-UK) over £20,000 in early-stage funding was distributed with hundreds of new ventures assessed using the system.

4.1 NVANA: New Venture Analytics

From a research perspective, the motivation for building such a system to support early-stage investment in new ventures two-fold: in order to i) gain greater insight into the investment process in practice ii) collect a suitable dataset from which to be able to assess the appropriateness of designing analytics. Figure 4.1 illustrates the intended role of the NVANA system.

NVANA intends to enable more efficient early-stage investment decisions. Specifically, it deals with the submission and assessment of business ideas ("Company") and will enable quick data input and subsequent evaluation of proposed ventures by a panel of experts ("Assessors") against a set of data and assessment criteria defined by an administrator ("Admin"). The outcome of such a system is a standardised set of data, scores and feedback ("Company Profiles").

The most recent version of NVANA was implemented with three different organisations UCL Advances, National Consortium of University Entrepreneurs (NACUE) and the China Innovation and Development Association in the UK (CIDA-UK) in running the following business plan competitions: UCL London Entrepreneurs' Challenge (E Challenge), CIDA-UK China UK Challenge and NACUE National Varsity Pitch Competition (NVPC).

NVANA's core is a self-serve customisable platform for reviewing business plan proposals (e.g., in the context of a business plan competition). The bulk of the functionality is used by the competition organiser (i.e., admin), while other sections are specifically for the expert reviewers (i.e., assessors) and applicants (i.e., companies). The organiser begins by defining their competition. This includes the types of businesses to be accepted (e.g., cleantech, mobile), types of applicants (.e.,g undergraduates, postgraduates, staff), and included regions and institutions. After completing the competition details, they enter the structure and dates of the competition's phases and are then ready to share the competition's application page. NVANA provides ongoing reports to the organiser on the progress with regards to applications and review by assessors.

Assessor feedback and scores are immediately compiled into a series of tables and charts which reduces data entry and increases the speed with which the various scores and opinions can be compared and acted upon. Feedback is available to be sent on to the applicants as soon as is desirable without any manual data entry. Finally, additional time is saved by automating miscellaneous tasks like assigning assessors to applications, normalising scores, and sending reminders.

Clearly there are differences from real-world investment by venture capital firms, or even angel investors, and business plan competitions. However, they served as a similar scenario, in which financial

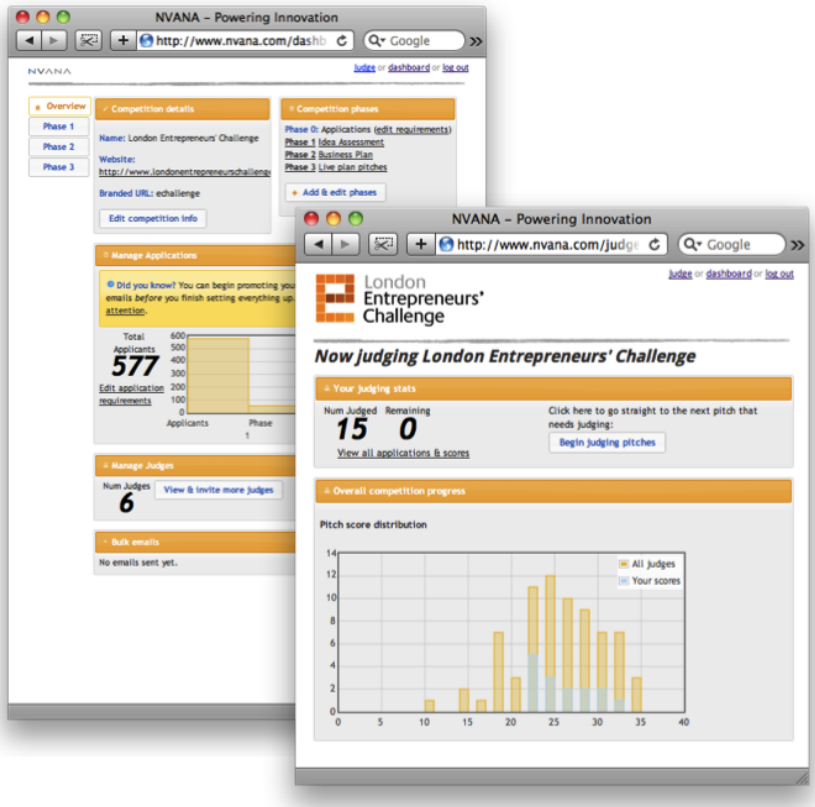


Figure 4.2: Screenshots of NVANA.

grants were distributed and therefore can be considered analogous to a real-world investment decision process. Obvious differences included assessment and decision by academics or external experts as opposed to professional investors and business plans often representing proposals for new ventures rather than actual existing businesses.

VentureRank

Beyond qualitative assessment by experts we had envisioned and implemented an experimental scoring system called VentureRank. The intention was to develop an algorithmic scoring or ranking of companies based upon the data input by a company and retrieving relevant external datasets on financials (e.g., Companies House), previous financings (e.g., CrunchBase), team members (e.g., LinkedIn), web traffic (e.g., Alexa) and search queries (e.g., Google Trends) amongst other various potential inputs. Figure 4.1 illustrates the inclusion of VentureRank in the NVANA system in order to provide both a qualitative assessment by experts and quantitative scoring algorithmically. The implementation of VentureRank is discussed in in Chapter 6.

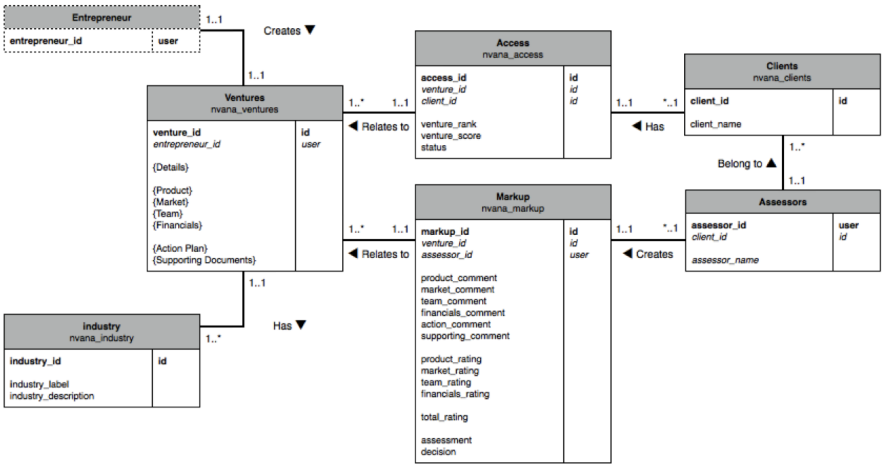


Figure 4.3: Entity Relationship (ER) Diagram for NVANA.

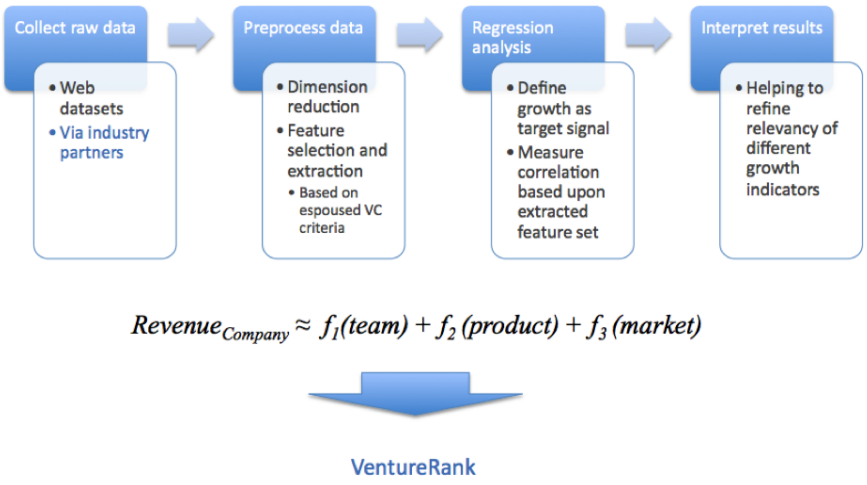


Figure 4.4: Proposed Process for Implementing VentureRank.

4.2 Limitations

Whilst NVANA removed some inefficiency in a typical investment workflow, largely in relation to data input and coordinating assessment, it did little to reduce the burden of assessing large number of business proposals. Several shortcomings were observed. Most importantly, the use of such as system must take into account the investment stage at which the data is collected in order to realistically obtain data worthy of any meaningful analysis. The companies' whose data was collected through previous implementations of NVANA rarely had sufficient operating track record. They were more often than not simply business ideas or plans. Therefore, whilst some relevant information was collected (e.g., team) others were not (e.g., financials). In addition, the data was highly qualitative and non-standardised across different implementations, leading to different silos of data.

In order to truly assess the viability of using computational tools in supporting early-stage investment there is scope for further development of the NVANA platform and concurrent implementation

with industry partners at the relevant stage of investment. Computational methods for early-stage investment has so far been subject to little academic inquiry, however, issues around quantitative analysis and standardisation must also be considered. There are clearly issues, unique to early-stage investment that may impact on the viability of using such techniques:

- Ability to quantitatively analyse an individual or team
- Different investment criteria used between investors
- Standardising all important aspects of a venture into an online questionnaire

However, additional research must be undertaken to see if these issues can be overcome and allow for successful implementation of computational analytics for enhancing early-stage investment.

Through our detailed survey of existing tools we have observe the nascent development of software and the limited application of analytical methods to support the venture capital investment process. Clearly adoption is still low and there is also a unanswered questions as to whether and how venture capital firms should best position themselves in order to take advantage of such new tools and capabilities (i.e., develop in-house versus selecting best-of-breed third party tools).

The following section and core of this research focuses on applying computational analytics (i.e., multi-label classification, recommender systems) to the domain of venture finance with the goal of improving company classification and applied to specific use cases (e.g., Identifying peers, Matching investors $\leftrightarrow$ companies). The next Chapter (see Chapter 5) covers improving upon existing industry classification schemes addressing the shortcomings of existing industry classification schemes (i.e., out-of-date, misrepresentation, misinterpretation). With advances in computational analytics and related techniques, it is possible to create an improved form of classification for describing the activities and relationships of private companies. Followed by (see Chapter 6 and 7), applying computational analytics to venture finance demonstrating the efficacy of recommender systems in relation to this novel application domain. Our methodology takes advantage of our access to venture financing data to improve the performance of recommendation models in improving private company similarity measures and the task of Top-N investment opportunity recommendation.

Part III

Analytics

Chapter 5

Multi-label Industry Classification

This chapter focuses on improving upon existing industry classification schemes. With advances in computational analytics and related techniques, it is possible to envision improved classifications that describe the activities and relationships of private companies. We put forward a methodology for generating a multi-label industry classification using supervised learning techniques. Then, we discuss our experimental results and possible extensions for future research.

The field of taxonomy and systematics deals with the question of *how* best to classify something. Researching taxonomies and classification schemes often surfaces notions such as phenetics and cladistics in natural sciences or perhaps the Dewey Decimal System [Dew76] in library and information sciences. These research areas of taxonomy and systematics are rich with literature and contributions which are clearly applicable to much broader domains than those mentioned above.

5.1 Private Company Classification

There is clearly scope to classify companies above and beyond simply a company's industry. In fact one of the most interesting and under-explored areas are the different possible dimensions of a classification scheme. Some of the potential dimensions and their origins are outlined in Table 5.1.

The study from which we adopt our model of venture capital decision process [TB84], for example, categorises the important factors for assessing a new venture as:

- *Market Attractiveness* — size, growth, and access to customers
- *Product Differentiation* — uniqueness, patents, technical edge, profit margin
- *Managerial Capabilities* — skills in marketing, management, finance and the references of the entrepreneur
- *Environmental Threat Resistance* — technology life cycle, barriers to competitive entry, insensitivity to business cycles and down-side risk protection

- *Cash-Out Potential* — future opportunities to realise capital gains by merger, acquisition or public offering

However, data required to classify against such dimensions is not always readily available (e.g., Managerial Capabilities) and often requires expert judgement or assessment. Preliminary experimentation in generating a multi-dimensional classification scheme baed upon the work of Osterwalder et al [OP10] in business model ontologies and their business model canvas is included in Appendix D.2.

Table 5.1: Proposed Additional Dimensions.

Proposed dimension	Example class labels	Source
Industry sector	Wholesale of computers, computer peripheral equipment and software Retail sale of computers, peripheral units and software in specialised stores Other software publishing	UK Standard Industry Classification (SIC) 2007
Market type	New Market Existing Market	[Bla09]
Customer type	Businesses Consumers Public sector	Startup Intelligence
Customer relationship	Personal assistance Dedicated personal assistance Self-service Automated services Communities Co-creation	[OP10]
Revenue model	Asset sale Usage fee Subscription fees Lending/Renting/Leasing Licensing Brokerage fees Advertising	[OP10]
Funding history	Venture Funding Total	CrunchBase, VentureSource
Location	Europe Western Europe United Kingdom London	AngelList
Stage	Discovery Validation Efficiency Scale Sustain Conservation	Startup Genome Compass

In relation to venture finance the most commonly used form of classification is industry classification schemes which attempt to organise companies by means of production or type of economic activity. However, even the term “industry” is not always well defined and such classification schemes may also relate to product offerings and behaviour in financial markets, amongst other factors.

5.2 Industry Classification

Whilst industry classification schemes have numerous uses (e.g., segmentation, comparison) they often have been designed by government bodies or agencies for the purpose of aggregating statistics. As noted by [BLO], in their study on capital market research, despite the widespread use of industry classification schemes by academic researchers, few studies directly test their efficacy. Several applications in

venture finance are reliant on some form of company classification, for example, screening prospective investment opportunities or identifying comparables for valuation purposes. In order to create an improved classification scheme, we must firstly evaluate the limitations of existing classification schemes. Secondly, we need to define what would make an improved classification scheme by looking at how to resolve such issues.

5.2.1 Existing Classification Schemes

Table 6.2 depicts some of the most common types of company classification schemes. Broadly these classification schemes fall into two main types: *public* and *proprietary*.

Table 5.2: Existing Public and Proprietary Company Classification Schemes.

Abbreviation	Full name	Sponsor	Criterion	
ISIC	International Standard Industrial Classification of All Economic Activities	United Nations Statistics Division	Production/Establishment	Public
NAICS[1]	North American Industry Classification System	Statistical bureaus of US, Canada, and Mexico	Production/Establishment	
NACE	Statistical Classification of Economic Activities in the European Community	European Community	Production/Establishment	
UKSIC	United Kingdom Standard Industrial Classification of Economic Activities	Office for National Statistics	Production/Establishment	
ICB	Industry Classification Benchmark	FTSE, Dow Jones	Market/Company	Proprietary
TRBC	Thomson Reuters Business Classification	Thompson Reuters	Market/Company	
RISC	Revere International Sector Classification	Revere Data	Market/Company	
VS	VentureSource	Dow Jones	Market/Company	
CB	CrunchBase	TechCrunch	Market/Company	
AL	AngelList	AngelList	Market/Company	

Public (or governmental) classification schemes are sponsored by national and supranational governments, agencies and other such political entities. These usually serve the purpose of helping provide governmental statistics such as employment and economic growth. For example, the United Kingdom Standard Industrial Classification of Economic Activities (UKSIC) which is sponsored by the Office for National Statistics.

Proprietary classifications schemes are those established usually by a private company for its own business activities and purposes. These usually serve the purpose of assisting individual companies business-to-business (B2B) marketing or sales activities. For example, the Industry Classification Benchmark (ICB), which was established by Dow Jones, an American publishing and financial information firm, and FTSE, a British provider of stock market indices and associated data services. Many proprietary classification schemes are closed (i.e., require a commercial license) however several are open particularly those used in open databases of technology companies. For example, CrunchBase which is maintained by the web-based technology publication TechCrunch (owned by AOL).

Whilst both public and proprietary classification schemes have different intended use cases and therefore classifying criterion it is useful to compare and contrast different aspects and characteristics of each. A potential source of ambiguity is the distinction between a company's "industry" and "market". This issue is largely addressed by [Mul03] who defines both: "a market consists of a group of current

and/or potential customers having the willingness and ability to buy products — goods or services — to satisfy a particular class of wants or needs. Thus, markets consist of buyers — people or organizations and their needs — not products"; and "an industry consists of sellers — typically organizations — which offer products or classes of products that are similar and close substitutes for one another".

For our purposes, due to restrictions of data access, we shall focus on the United Kingdom Standard Industrial Classification of Economic Activities (UKSIC), VentureSource, CrunchBase and AngelList classification schemes.

United Kingdom Standard Industrial Classification of Economic Activities (UKSIC)

A Standard Industrial Classification (SIC) was first introduced into the UK in 1948 for use in classifying business establishments and other statistical units by the type of economic activity in which they are engaged (Office for National Statistics 2012). According to the Office for National Statistics [Off07] the classification provides "a framework for the collection, tabulation, presentation and analysis of data, and its use promotes uniformity [which] can be used for administrative purposes and by non-government bodies as a convenient way of classifying industrial activities into a common structure". The classification has been revised periodically in 1958, 1968, 1980, 1992, 1997, 2003 and, most recently, in 2007. Issues with the UKSIC system can be demonstrated both from relevant literature and empirical analysis showing that UKSIC codes are not fit for purpose. An industry-led report focused on the funding of technology in Britain [GMPR07] noted the following failings: "analysis is beset by measurement and definition issues (not only are SIC codes a blunt instrument but delivery mechanisms may have a perverse effect as well — software delivered on a disc will count as manufacturing but software delivered over the internet will not", furthermore, "the overall size of the technology market is hard to quantify (reliance on SIC codes alone has proved ineffective)".

Moreover, data provided by Bureau van Dijk FAME database also shows failings of UKSIC classification. Observing a sample of 30,221 active companies based in the London WC postcode area it is possible to observe the different UKSIC classification code for each company, of which there are 517 different UKSIC (2003 revision) classifications in total. From our sample of companies in the WC postcode area 353 unique UKSIC codes were found but more importantly around one third of companies (33.94%) were unclassified and another third (31.61%) were classified as "7487" which refers to "Other Business Activities". Both the primary and secondary [GMPR07, NR13] reports of using UKSIC codes as a means for classification have identified clear limitations:

- *Often out-of-date and revised intermittently* — classification does not cover new markets, industries and business models which are indicative of innovation high-growth technology companies
- *Self-reported classification* — leading to scope for misinterpretation and misrepresentation as

demonstrated by the large number of companies classified as "Other Business Activities"

VentureSource

Figure 5.1: Network Graph Showing Industry Hierarchy of VentureSource.

The VentureSource classification scheme is similar to various governmental schemes such as Standard Industry Classification (SIC) codes in that it provides a hierarchical, fully assigned, industry-centric classification. VentureSource offers three levels of hierarchy for industry classification: industry groups (1st level); industry segments (2nd level); and industry codes (3rd level). Figure 5.1 depicts the three tiers (Group labeled in large blue font, Segment labeled in light blue font, and Code the unlabelled nodes). Group describes broad industry sectors (e.g., "Information Technology") and the subsequent tiers Segment (e.g., "Software") and Code (e.g., "Communications Software") provide further granularity. This industry classification is similar to other public (e.g., UKSIC, USSIC, NAICS) or private (e.g., Capital IQ) classification schemes. It offers a more sophisticated classification scheme than other similar datasets such as CrunchBase which uses a simple category code to depict industry sectors (e.g., "Web", "Mobile"). Around 1,500 or 6.8% of companies in VentureSource either have no industry Code or are classified as "To be assigned". In terms of limitations of the VentureSource classification scheme the most obvious is the use of *Full assignment*. Full assignment (i.e., exclusive class membership) is where a company can only be defined as a single class, such as "Information Technology" or "Healthcare", but but not both (i.e., multiple assignment) or a mix of two or more classes (i.e., partial assignment). We

believe this form of class assignment is overly simplistic and leads to reduced utility.

CrunchBase

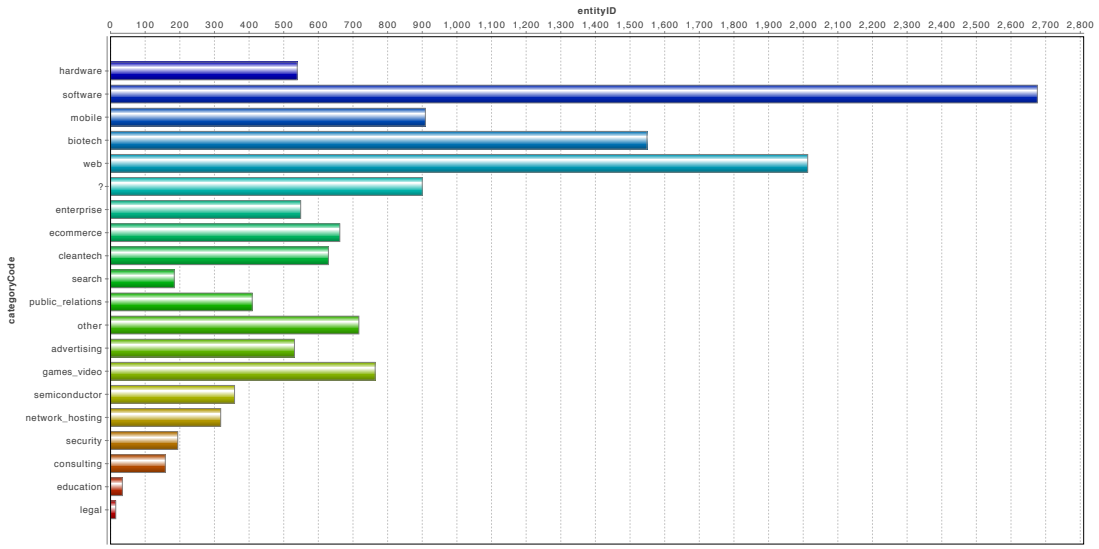


Figure 5.2: Category Code Distribution of CrunchBase.

CrunchBase can be best viewed as a "repository" of start-up companies, individuals, and investors with a focus on US high-tech sectors [ABSW11]. CrunchBase is operated by TechCrunch (owned by AOL), a popular Internet blog largely focused on technology and Internet companies. As noted by others [ABSW11] the companies covered by CrunchBase span a wide spectrum from large multi-billion dollar businesses, like Google or eBay, to small privately held companies with very few employees only recently founded. CrunchBase provides an industry category code used to depict broad industry sectors mainly related to technology companies (e.g., "Games, Video & Entertainment", "Mobile"). This industry classification is quite simplistic compared to other public (e.g., USSIC, NAICS) or private (e.g., Dow Jones VentureSource) classification schemes which often have an industry hierarchy with several tiers for broad industry sectors (e.g., "Information Technology") and subsequent sub-sectors (e.g., "Software") providing further granularity. Around 900 or 6.4% of companies in CrunchBase have no assigned category code and are currently unclassified. Limitations of the CrunchBase categories as classification scheme include *Poor class labels* including over represented and catchall classes (e.g., "Software", "Web") which could arguably describe the vast majority of companies covered by CrunchBase database. In fact the category codes "Software" and "Web" account for 19% and 14% of all companies in our dataset respectively. In comparison to VentureSource with has 200+ low level industry codes to describe companies CrunchBase has less than 20 category codes.

AngelList

AngelList, founded in 2010 allows entrepreneurs to post their funding needs to a self-described "platform" [Ang13a] for start-ups to raise funding, which in turn is reviewed by investors. The service was recently given support by the Kauffman Foundation and has seen wide-spread adoption in industry, having connected over 1,500 new companies with thousands of angel investors and venture capital firms [Rav12]. In terms of class labels, AngelList boasts the "most accurate set of market names in the world" which have been independently self-reported by entrepreneurs and investors. The classification scheme employed also has a interesting structure in using a direct acyclic graph to order the various classes or "Markets" as they are referred to by AngelList.

Limitations of the AngelList classification scheme:

- *Unidimensional* — the scheme does not distinguish between different aspects of a company's activities and relationships, for example the following tags (used by AngelList) can potentially relate to the different aspects of a company: "Enterprise Software" (i.e., Market), "Finance" (i.e., Industry), "Consumer", (i.e., Customer type), "SaaS" (i.e., Business model). Without this meta-information the same class tags potential could represent quite different companies. For example using the tags "Software" and "Finance" there is no knowledge of customer type (i.e., Consumer, Enterprise) which leads to ambiguity and therefore could represent two quite different companies. Using real-world examples, MoneyDashboard (a consumer focused personal banking software) and OpenGamma (an enterprise focused risk management software for hedge funds and investment banks) could both be described with these two tags.

- *Limited Coverage* — largely due to its relative infancy and focus on being a "platform" over primarily a database or data provider AngelList has less coverage than other alternatives such as CrunchBase and VentureSource. As with any user generated system often companies have incomplete information (e.g., "market tags").

5.2.2 Issues with Existing Schemes

The issues previously discussed are further illustrated by the following classification of three companies: Duedil Ltd. is a London-based technology company self-described as "a one-stop shop for business information and intelligence" [2]; Apsalar Inc. describes itself as a provider of "mobile app analytics and advertising solutions"[3]; finally, FrameHawk Inc. also based in San Francisco, provides "application mobilization"[4]. Table 5.3 below shows how three different companies Duedil Ltd., Apsalar Inc., FrameHawk Inc. are classified by the different classification schemes.

The limitations of the above scheme are self-evident with UKSIC proving either out-of-date or

[2] Duedil Ltd. — `https://www.duedil.com` [3] Apsalar Inc. — `https://apsalar.com` [4] FrameHawk Inc. — `http://www.framehawk.com`

Table 5.3: Example Classification of Duedil Ltd., Apsalar Inc. and FrameHawk Inc.

Company	UKSIC	VentureSource	CrunchBase	AngelList
Duedil Ltd.	"7487 — Other business activity"	-	"Enterprise"	"Financial Services", "B2B", "Marketplaces", "Big Data"
Apsalar Inc.	"7414 — Business Consultancy"	"Information Technology" "Software" "Vertical Market Applications Software"	"Advertising"	-
FrameHawk Inc.	-	"Information Technology" "Software" "Vertical Market Applications Software"	"Software"	-

misinterpreted; VentureSource[5] fully assigned classes (i.e., either Information Technology *or* Business and Financial Services but not both); CrunchBase proving overly simplistic with limited number of available classes; and AngelList giving a richer by still ambiguous, and often non existent, classification of the company in question. To summarise the limitations of the different existing classification schemes:

- *Out-of-date and revised intermittently* — this is a key issue for public governmental schemes, for example the latest revision of UKSIC was in 2007. The scheme therefore fails to address high-growth technology companies

- *Self-reported* — leading to scope for misinterpretation of a classification scheme and misrepresentation of how a company should be classified

- *Full class assignment* — companies are fully assigned to either a single class or multiple classes but cannot be partially assigned to different classes (i.e., multi-label classification)

- *Unidimensional* — leading to ambiguity around how class labels relate to the different aspects (or dimensions) of a company, its relationships and activities (i.e., Market, Industry, Customer, Product, etc)

- *Limited Coverage* — no classification scheme has 100% coverage of all companies but there is also a large amount of missing data or unknown industry classifications accounting for 6.8% of VentureSource and 6.4% of CrunchBase respectively

In designing an improved classification scheme there are several important considerations to be made. At this point it is useful to revisit the existing classification schemes and the key attributes or characteristics of the different schemes (i.e., Class labels, Dimensions, Structure and Class assignment) as shown in Table 5.4.

The different characteristics of a classification scheme are clearly important considerations in attempting to design an improved scheme. In order to clarify the different characteristic we shall define each:

[5] VentureSource industry classification has three tiers: Group, Segment and Code

Table 5.4: Characteristics of Existing Classification Schemes.

	Classification Scheme			
Characteristic/Property	UKSIC	CrunchBase	VentureSource	AngelList
Class labels	"63120 — Web portals"	"Web"	Group "Information Technology" Segment "Software" Code "Communications Software"	"Information Technology" "Software" "Enterprise Software"
Dimensions	Industry	Category	Industry	Market Location
Structure	Tree	Bins	Tree	Graph
Assignment	Full (Primary), Multiple (Secondary)	Full	Full	Multiple

- *Class labels* — serve to classify the company into one of a number of distinct categories. In designating class labels it is possible to use existing class labels from existing classification schemes, or emergent class labels, either manually defined or through unsupervised learning techniques.

- *Dimensions* — the dimensions (also referred to as "aspects" [AA12]) used to classify a company, its activities and relationships (e.g., Industry, Customer type, Location)

- *Structure* — the structure of a classification scheme dictates the relationship and inheritance between different classes. Different classification structures include: *Bins* (or buckets) of independent classes (e.g., CrunchBase); *Tree* (or hierarchical) structure with a hierarchy of classes with parent and child classes (e.g., UKSIC); *Tags*, a non-hierarchical keyword or term assigned to a piece of information; and *Graph* structure of classes (i.e., nodes or vertices) where some pairs of the classes are connected by links (i.e., edges) (e.g., AngelList).

- *Class Assignment* — how companies are assigned or attributed to different classes. Different variations include: *Full* assignment whereby a company is fully assigned membership exclusively to a single class; *Multiple* assignment, companies are fully assigned membership to multiple classes; and *Partial*, companies are partially assigned membership to multiple classes.

Our specific focus is around industry assignment or how companies are assigned to different industry classes (or categories). An inherent limitation of existing industry classification schemes, means companies must be fully assigned to a single industry class (at each tier of the industry hierarchy). There is no notion by which a company may be assigned to more than one single class (i.e., multiple assignment) or in different proportions across classes (i.e., partial assignment). This is a common limitation amongst several widely adopted industry classification schemes (e.g., CrunchBase, VentureSource).

In order to generate multiple industry assignment for each investee private company, we propose a supervised learning approach whereby we are given a description of a document and a fixed set of classes (or labels) (see Section 5.4).

5.2.3 Resolving Issues with Existing Classification Schemes

As stated at the beginning of this chapter, as well as reviewing existing schemes (hereafter, referred to as $I_{Existing}$) we need to define what would make an improved classification scheme by looking at how to resolve the issues with the current schemes. In terms of how the limitations might be addressed by an improved classification scheme. There is potential to focus on a dynamic scheme addressing high-growth technology companies with up-to-date class labels rather than out-of-date and revised intermittently (e.g., UKSIC 2007 is latest revision). Firstly, we focus on auto-classification using textual data from companies websites to resolve incomplete or non-existent classifications, furthermore, the issue of companies self-reporting classifications leading to scope for misinterpretation and misrepresentation. Secondly, we tackle full assignment (i.e., exclusive class membership) which doesn't reflect companies in the real world. We focus on multi-label classification through partial (or multiple) assignment of classes.

Whilst there is potential for improvement across several different characteristics we have focused on automating classification of class labels and generating a multi-label classification scheme (i.e., class assignment). Therefore, the two major areas of improvement we have chosen to focus upon will lead to an automated and multi-labeled classification scheme:

- **Automated** — auto-classification of company's industry (referred to as I_{Auto}) using textual data from company website we reduce the scope for misinterpretation of the scheme and misrepresentation of an individual company (see Section 5.3)

- **Multi-labeled** — a multi-labeled scheme classifying a company beyond purely a single industry or sub-industry classification using multiple (referred to as $I_{Multiple}$) and partial (referred to as $I_{Partial}$) assignment (see Section 5.4)

Other potential improvements and characteristics (e.g., Structure, Dimensions) not directly covered are also discussed (see Section 5.6).

5.3 Supervised Learning Using Existing Classification Schemes

Our goal is to auto-classify using textual data from company description (i.e., from existing databases) or company websites. We propose the use of textual data due to its relative accessibility (i.e., from open databases such as CrunchBase or using web scraping tools) and universality (as more and more companies have a presence online). Broadly there are two type of approaches to this form of classification problem: *Supervised Learning* approaches using training data to infer model and then applying model to test data (e.g., Naive Bayes, Ensemble classifiers); and *Unsupervised Learning* approaches without training data both model inference and application rely on test data exclusively (e.g., Clustering, Factor analysis, Topic modeling). An exploratory analysis of some of the many different unsupervised

techniques has been undertaken. Whilst some factor analysis techniques (e.g., PCA) are used in pre-processing we utilise a supervised learning methodology.

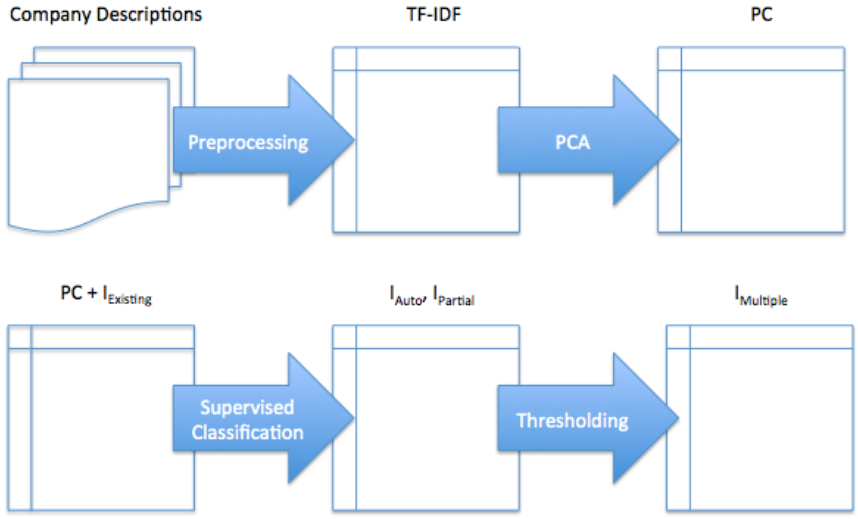


Figure 5.3: Process Diagram for Auto Classification and Multi-label Classification.

As shown in Figure 5.3 our overall process for auto classification against existing schemes and subsequent multi-label classification is as follows:

- Prepare input data of companies with textual descriptions and existing industry classification scheme (e.g., VentureSource) (see Section 5.3.1).
- Preprocess company textual descriptions (stemming, n-grams) and generate term-frequency inverse document frequency (TF-IDF) vectors (see Section 5.3.2) and principal component analysis (PCA) on TF-IDF vector to reduce dimensionality.
- Learn a supervised classification model (see Section 5.3.4) using various classification algorithms (see Section 5.3.5) using our principal components as input feature vector and industry classification as label.
- Use output as industry classification I_{Auto} and also as input for multi-label classification (see Section 5.4.

5.3.1 Datasets

Working alongside industry partner Correlation Ventures, a US venture capital firm, we have been granted access to data provided by Dow Jones VentureSource, a data provider for the venture capital industry. The VentureSource dataset includes industry hierarchy classification for companies related to historical venture financings in the US since 1987. We also have a principal component representation of the investee private companies' descriptions. Further data was accessed via publicly available application programming interfaces (API) for both CrunchBase and AngelList. This allowed us to consider

three main classification schemes employed by CrunchBase (i.e,. Category), AngelList (i.e., Markets) and VentureSource (i.e., Industry).

The current approach to auto-classification has been a typical supervised approach based upon an existing classification schemes (i.e., CrunchBase, VentureSource and AngelList) and some textual description. Whilst we have access to several different textual descriptions (e.g., Company website, CrunchBase overview, VentureSource company description and AngelList product description) we have opted to use the VentureSource company descriptions (i.e., due to privacy restrictions) and CrunchBase overview.

In order to maintain anonymity of the dataset provided by Correlation Ventures and Dow Jones VentureSource our data includes no obvious identifiers (e.g., Company name, Website, Description). Instead we simply have been provided with principal components derived from a singular value decomposition (SVD) of VentureSource company description. In the case of CrunchBase (also AngelList) we require a preprocessing step in order to use the company description (e.g., overview) or textual data retrieved from the company website.

5.3.2 Textual Preprocessing

Given a set of documents D with company descriptions d in our case textual descriptions of the companies. For example:

$\langle d \rangle = \langle$*Apsalar provides app developers and publishers with Mobile Engagement Management (MEM) solutions that increase user engagement, retention, and monetization. By leveraging Apsalar's best-in-class mobile analytics and integrated behavioral targeting capabilities, customers can analyze, optimize, and monetize user engagement in their apps. Apsalar enables engagement with the same user anytime, anywhere, in a publisher's apps or in 3rd party apps, for any targeted promotion. Founded in 2010 and based in San Francisco, Apsalar is backed by leading venture investors, Thomvest Ventures, Battery Ventures, and DN Capital. Apsalar has won numerous awards, including eWeek's Top Ten Promising Mobile IT Startups in 2011.*$\rangle$.

The following preprocessing steps were used on our textual descriptions as follows. If using text from companies website we all require removing HTML tags (e.g., $\langle p \rangle$,$\langle /p \rangle$) and other markup.

- *Transform Cases* – transforms cases of characters in a document to lower case
- *Tokenize* – splits the text of a document into a sequence of tokens using non-letter characters as splitting point

- *Filter Stopwords* – removes English stopwords from document
- *Stem (Porter)* – stems English words using the Porter stemming algorithm applying an iterative, rule-based replacement of word suffixes intending to reduce the length of the words until a minimum length is reached
- *Generate n-Grams (Terms)* – create term 3-grams of tokens in a document
- *Filter Tokens (Length)* – filters tokens based on a minimum length of 4 characters

We use the above steps in processing the documents (i.e., company descriptions) and generate word vectors in term-frequency inverse document frequency (TF-IDF) form.

5.3.3 Dimensionality Reduction

For the purpose of our supervised classification we generate a term-document matrix [MRS08]: an $m \times n$ matrix C, each of whose rows represents a term and each of whose columns represents a document in the collection. Even for our of modest size collection of documents $n = 20,000$, the resulting term-document matrix C has several tens of thousands of rows and columns. In order to efficiently classify such a collection we decide to use a dimension reduction technique though matrix decomposition called singular value decomposition (SVD). Specifically, we utilise principal component analysis (PCA) and a common method for implementing PCA is by applying SVD to the covariance matrix. Singular value decomposition (SVD) and principal component analysis (PCA) are both eigenvalue methods used for dimension reduction of a high-dimensional dataset (e.g., text, images, etc) whilst retaining key information.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. When using a high-dimensional dataset (e.g., textual data) one commonly used technique is PCA an eigenvalue based factor analysis method. First developed by Pearson (1901) it transforms a number of positively correlated variables into a smaller number of uncorrelated variables or components. PCA is a variance-focused approach and seeks a linear combination of variables with the maximum variance extracted from the original variables, resulting in the first component and then recursively creates additional components. Therefore the first component accounts for as much variance as possible and succeeding components account for a decreasing amount of variability. A more detailed comparison and discussion of SVD and PCA is covered in the literature [Alb04].

Figure 5.4 illustrates how PCA can lead to a useful delineation of the different companies and by only using the first two components (i.e., those that account for the most variance) we have started to separate out companies by different categories as illustrated. Once we have the principal components

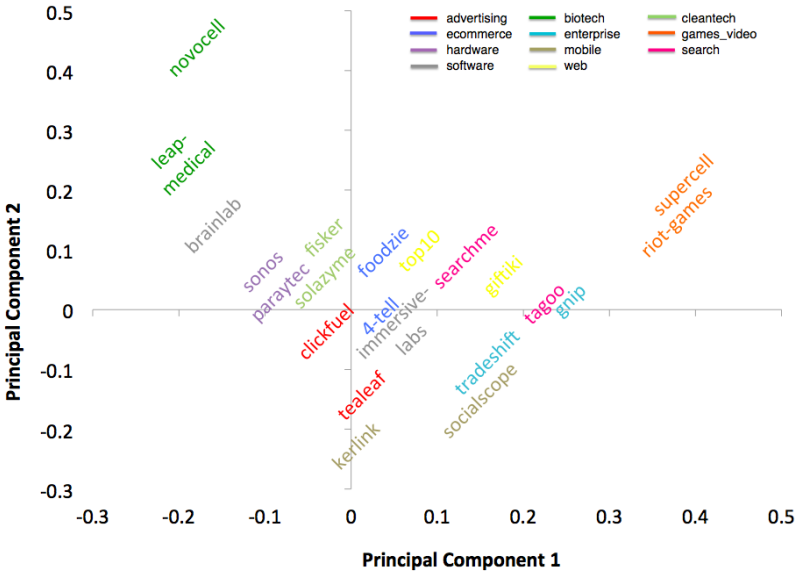


Figure 5.4: Companies Plotted Against First and Second Principal Components with Colour Representing CrunchBase Category.

derived from a textual description plus the relevant classification schemes we can continue with our supervised classification using this information as our training set.

5.3.4 Supervised Learning

Using a sample of around 20,000 private US companies we trained various supervised classifiers to predict various tiers of the VentureSource industry classification scheme (i.e., Group, Segment, Code) as output and principal components derived from a singular value decomposition of VentureSource company description as our input feature vector.

In order to generate our multi-label industry vectors we propose a supervised learning approach whereby we are given a description $d \in X$ of a document, where X is the document space, and a fixed set of classes (or labels) $C = \{c_1, c_2, \ldots, c_j\}$ [MRS08]. In our case classes represent the industry categories and documents textual descriptions of the companies. We are given a training set D of labeled documents $\langle d, c \rangle \in X \times C$. For example:

$$\langle d, c \rangle = \langle \text{Digg is a user driven social content website,} \\ \text{Information Technology} \rangle.$$

Using a learning method or algorithm, we then wish to learn a classifier that maps documents to classes:

$$\gamma : X \to C$$

We denote the supervised learning method by Γ and write $\Gamma(D) = \gamma$. The learning method Γ takes the training set D as input and returns the learned classification function γ. This type of learning is called supervised learning because a supervisor (the human who defines the classes and labels training documents) serves as a teacher directing the learning process.

Through implementing various learning methods (Naïve Bayes[6], SVM, Random Forest) we learn a classification function γ for each industry class label c (i.e., binary classification) for all companies with textual descriptions. This process allows us to classify new companies against an existing scheme (e.g., VentureSource, CrunchBase).

5.3.5 Classification algorithms

A number of different learning methods we implemented and compared for our supervised learning classification. Figure 5.5 depicts some characteristics of different learning methods [FHT08]. The experimental results for various classification algorithms are included in Section 5.5.3.

Some characteristics of different learning methods. Key: ▲= good, ◆=fair, and ▼=poor.

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of "mixed" type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

Figure 5.5: Characteristics of Different Learning Methods [FHT08].

Naïve Bayes

Naïve Bayes is a probabilistic classifier based on applying Bayes' theorem. For example, if we suppose documents are drawn from a number of classes of documents which can be modelled as sets of words where the (independent) probability that the ith word w_i of a given document occurs in a document from class C can be written as:

[6] Naïve Bayes offered superior classification accuracy.

$$p(w_i|C)$$

Then the probability that a given document D contains all of the words w_i given a class C is:

$$p(D|C) = p(w_i|C)$$

It is "naïve" in the sense that, it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. For example, a company may be considered to belong to the industry class "Information Technology" if it's description has a high occurrence of the terms "Information" and "Technology". Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this company belongs to the "Information Technology" industry.

Beyond Naïve Bayes classification we also implemented several other supervised classifiers for comparison. For more detailed information and mathematical formulations see Appendix E.

Decision Tree (c4.5)

c4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. The general algorithm for building decision trees:

- Check for base cases
- for each attribute a find the normalised information gain from splitting on a
- let a_{best} be the attribute with the highest normalised information gain
- create a decision node that splits on a_{best}
- recurse on the sublists obtained by splitting on a_{max} and add those nodes as children of node

Information gain is based on the concept of entropy used in information theory.

Random Forests

Random Forests are a class of ensemble classifiers developed by Leo Breiman and Adele Cutler. Our implementation used developed under Weka [7] software by University of Waikato. Each tree is constructed using the following algorithm:

[7] Weka — `http://www.cs.waikato.ac.nz/ml/weka/`

- Let the number of training cases be N and the number of variables in the classifier be M where $N \geq M$
- choose a bootstrap sample (i.e., training set for this tree by choosing n times with replacement from all available training cases)
- use the rest of the cases to estimate the error of the tree, by predicting their classes
- for each node of the tree, randomly choose m variables on which to base the decision at that node
- calculate the best split based on these m variables in the training set
- each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier)

Random Forests consist of many decision trees and outputs the class that is the mode of the classes output by individual trees.

Support Vector Machines (SVM)

Support Vector Machines (SVM) are a non-probabilistic binary classifier developed by Vladimir Vapnik and Corinna Cortes. Our Implementation used developed under LibSVM library as formally defined by Chih-Chung Chang and Chih-Jen Lin [CL11].

Generally SVM classifies as follows:

- Given a set of training examples, each marked as belonging to one of two classes, an SVM training algorithm builds a model that assigns new examples into one category or the other
- An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap (i.e., the hyperplane) that is as wide as possible
- New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on

5.4 Multi-label Classification

This section describes the methodology of using the confidence interval from supervised learning (see Section 5.3.4) based upon existing schemes outlined previously (e.g., VentureSource, CrunchBase) in order to create a multi-labeled classification scheme through partial or multiple assignment.

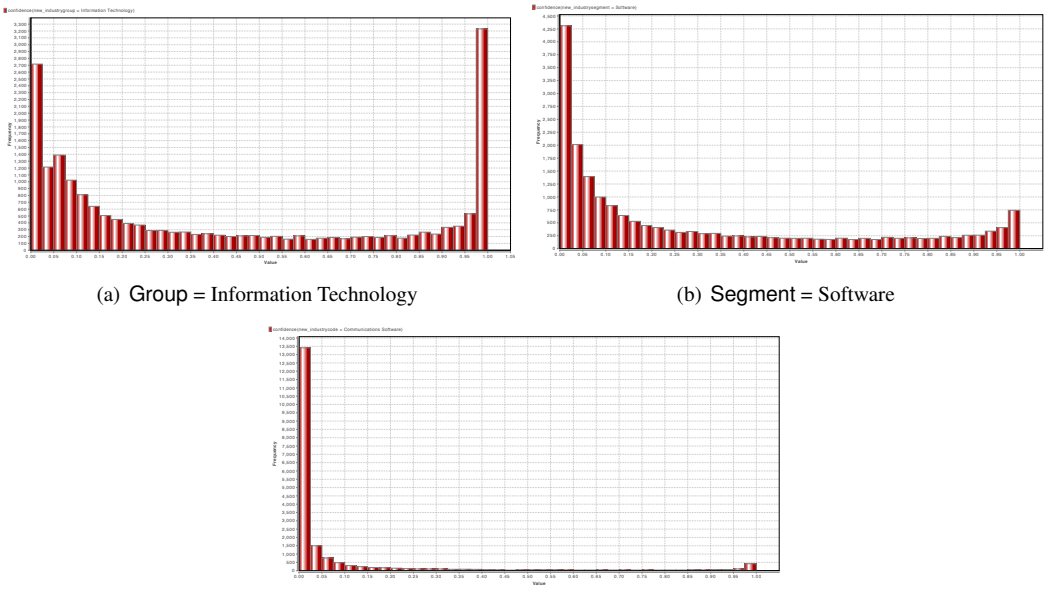
We define our existing industry classification schemes (e.g., VentureSource, CrunchBase) as $I_{Existing}$. Through implementing various learning methods (see Section 5.3.4) we learn a classification function for each industry class for all investee private companies with textual descriptions. This process allows us to auto classify new companies against an existing scheme I_{Auto} but also to generate novel class assignments (i.e., multiple, partial). By using the confidence level output of the classifier γ

for each class c we can associate a partial class membership in order to generate $I_{Partial}$. Furthermore, by simply defining a threshold confidence level (e.g., 0.5) we can generate $I_{Multiple}$ which is similar to categorical "tags" used in several other domains. For our purposes we define three distinct forms of class assignment where I describes our industry vector for a company with industry classes c:

$$
\begin{aligned}
I_{Existing} &\approx c_1, c_2, \ldots, c_n \;\; where\; 0 \leq c \leq 1 \in \mathbb{Z}\; and \sum_{1}^{n} c_j = 1 \\
I_{Auto} &\approx c_1, c_2, \ldots, c_n \;\; where\; 0 \leq c \leq 1 \in \mathbb{Z}\; and \sum_{1}^{n} c_j = 1 \\
I_{Multiple} &\approx c_1, c_2, \ldots, c_n \;\; where\; 0 \leq c \leq 1 \in \mathbb{Z}\; and \sum_{1}^{n} c_j > 1 \\
I_{Partial} &\approx c_1, c_2, \ldots, c_n \;\; where\; 0 \leq c \leq 1 \in \mathbb{R}\; and \sum_{1}^{n} c_j > 1
\end{aligned}
\tag{5.1}
$$

In terms of class membership we distinguish between partial (i.e., non-uniform) and multiple (i.e., uniform) assignment. It is important to note that $I_{Existing}$ and I_{Auto} take the same form in terms of class assignment (i.e., whereby companies are *fully* assigned to a single class) but I_{Auto} will result in no missing or unclassified companies so long as they have a textual description and potentially reclassification of certain companies based upon the outcomes of the supervised learning methodology.

As noted by [WVSB11] several application domains exist in which multi-class or multi-labeled classification problems arise whereby instances do not simply belong to one particular class, but exhibit a partial membership to several classes. In fact, multi-label learning has received significant attention and resulted in the development of a variety of multi-label learning methods [HVS+12].



(a) Group = Information Technology

(b) Segment = Software

(c) Code = Communications Software

Figure 5.6: Example Distribution of Confidence Levels for Binary Class Membership of Group, Segment and Code.

5.4.1 Partial Class Assignment Using Confidence Intervals

Partial assignment refers to the concept of allowing a particular entity to be classified to more than one class in a non-uniform manner. Partial assignment (e.g., Company A is 30% "Enterprise" and 70% "Software") may also be referred to weighted or probabilistic assignment. It differs from multiple assignment in that classes are not equally or uniformly assigned. Using the output of the previous auto-classification against existing schemes (see Section 5.3.4) it is possible to retrieve a confidence interval for each class defined in the training set data. For partial assignment we can simply use this confidence interval as the partial assignment value (i.e., weighting, probability) for class membership.

5.4.2 Multiple Class Assignment Using Thresholding

Multiple assignment refers to the concept of allowing a particular entity or instance to be classified uniformly to more than one class (e.g., Company A belongs to both "Enterprise" and "Software" classes). In terms of implementing multiple assignment we can again utilise the confidence intervals output from previous supervised classification against existing schemes. However, instead of simply using the confidence interval we define a threshold above which an entity has class membership by generating a binomial representation of the companies classification. Clearly how we define this threshold level for class membership is important, initially we have chosen to use the quotient of one divided by the total number of classes. This will allow assign only classes which have an above average confidence interval.

5.5 Experiments

5.5.1 Experimental Settings

Using supervised learning techniques to classify private companies against existing industry classification schemes. We will then derive standard performance metrics based upon cross validation. We have tested our supervised learning and multi-label classification methodologies using both the VentureSource and CrunchBase datasets. Below we outlined the measures and results of our experiments.

For the supervised classification methods we use RapidMiner by Rapid-I[8] an open-source rapid prototyping environment for knowledge discovery and data mining. Previously known as YALE (Yet Another Learning Environment) the project was developed by the Artificial Intelligence Unit of the Dortmund University of Technology (Klinkenberg, Mierswa Fischer 2001).

For multi-label classification we define a threshold level and assign classes accordingly.

5.5.2 Experimental Measures

In order to evaluate the accuracy of the supervised classification we perform a ten-fold cross validation (X-validation) and report average % performance metrics. The input example set is split up into number of validations subsets using stratified sampling which builds random subsets and ensures that the class

[8] RapidMiner — `http://rapid-i.com`

distribution in the subsets is the same as in the whole example set. These subsets are then used to train and test our model by applying the trained model and measuring its performance (i.e., Accuracy, Precision, Recall). Confusion matrices for each classifier with class precision and class recall values are included in Appendix C.

5.5.3 Experimental Results

VentureSource

Table 5.5: Performance of Auto Classification on VentureSource Dataset.

	Accuracy		
Technique	Group	Segment	Code
Decision Tree (c4.5)	56.84%	26.12%	7.57%
W-RandomForest	71.39%	54.32%	26.26%
libSVM	68.11%	53.92%	25.95%
Naïve Bayes	61.90%	45.87%	25.26%

The main observations were as follows: i) W-RandomForest classifier achieved high accuracy across all three tiers of the hierarchy ii) Naïve Bayes benefits from being much faster than other techniques (e.g., libSVM, W-RandomForest) iii) Accuracy deteriorates for classification of lower tiers (textsf, Code) with increased number of classes. In order to improve the classification accuracy we use binary classification of each industry class. For the detailed experimental results for each binary classifier please refer to Appendix C. A potential alternative to increase classification accuracy further is described under extensions and improvements (see Section 5.6.1).

CrunchBase

Table 5.6: Performance of Auto Classification on CrunchBase Dataset.

Technique	Accuracy	Time (ms)
Decision Tree (c4.5)	37.95%	2077
W-RandomForest	45.51%	14123
libSVM	20.26%	23153
Naïve Bayes	46.16%	28

The main observations were as follows: i) Naïve Bayes classifier achieved increased accuracy over other classifiers ii) Incremental improvements from additional preprocessing steps (e.g., N-grams, Stemming, Normalising) iii) Cross-validation achieving high accuracy on binary classification, much less accurate on all classes (i.e., multi-class classification) iv) Naïve Bayes classifier had the fastest training time of only 28ms on average

Even with perfect even oracle-like accuracy we would be auto-classifying companies using an imperfect classification schemes for the reasons mentioned previously (flat structure, full assignment, poor class labels, etc). If we observe the distribution of companies in each class from our CrunchBase training (see Figure 5.2) set we see two dominant classes, namely “Web” and “Software”. This overrepresentation of certain classes, especially led to overfitting and misclassification due to the dominance of these classes.

5.6 Discussion

We have reviewed existing classification schemes and the various limitations of such schemes. Subsequently, we have proposed an alternative in multi-label industry classification through a supervised learning methodology. In terms of specific improvements of our generated classification schemes:

- Auto classification (I_{Auto}) against existing classification schemes helps to address the issues of misinterpretation (i.e., of classification scheme) and misrepresentation (i.e., of individual companies). By using a supervised learning technique versus manual or self-reported classification we remove any risk of human error in misinterpreting the classification scheme or misrepresenting an individual company. The classification is based purely on the textual description and the training set of classified companies.

- Furthermore I_{Auto} deals with the problem of incomplete coverage in existing schemes accounting for around 6% in both CrunchBase and VentureSource datasets unclassified or defined as "To be assigned". This is a small but not insignificant proportion of all companies.

- $I_{Partial}$'s assignment allows a consideration of *similarity* which is simply not possible (or extremely limited) with full assignment whereby companies are assigned to a single industry class label.

- $I_{Multiple}$'s assignment allows companies to have multiple industry "tags" which will be utilised later as an input for refining our recommendation techniques (see Chapter 6 and 7).

Arguably using the principal components (i.e., PC) as our measure of similarity is a simpler alternative to generating a novel industry classification such as $I_{Partial}$. Whilst we have improved upon existing classification schemes through I_{Auto} we still do not have a methodology to introduce new class labels in order to represent new industries. Two areas of potential extension and improvement yet to be discussed in detail are the *structure* and *dimensions* of the classification scheme.

5.6.1 Hierarchical Classification

In terms of further extensions we have also considered the possibility of hierarchical classification. Traditionally classification schemes used to defined the relationships and activities of businesses or companies have been confined to strict hierarchies and rarely subject to other types or structures such as tags or graphs. However, in the case where classification schemes have some structure (e.g., Tree/Hierarchical, Graph) such as with VentureSource it is possible to use industry hierarchy information in supervised classification to improve classification (e.g., use industry Group to classify industry Segment). Initial experiments indicated an increase in overall classification accuracy, however, if it difficult to know what extent we are reducing noise versus biasing our classification.

As a simple illustration, using principal components to train a binary classifier to evaluate whether 100 entities belong to industry segment "Software". Without using the industry group as an input our classifier split test accuracy is 79%. Using the industry group as an additional input feature our accuracy increases to 88%. However, presume an individual company description included occurrences of the term "program". Is this simply reducing noise (i.e., description was actually "healthcare program" in another context outside of software therefore is rightly given a lower confidence) or biasing the classification (i.e., company description included the term "program" in relation to software but is classified under the industry group "Healthcare" but is given lower confidence). An alternative approach would be to input the industry Group confidence scores vector (i.e., not the binary indicator).

5.6.2 Use Cases in Venture Finance

We have demonstrated and detailed the methodology for generating a novel form of industry classification and have shown how it overcomes several of the issues of existing classification schemes. Through utilising supervised learning we diminish the joint issues of misrepresentation and misinterpretation. By using company websites as the source of our textual input data we could also eliminate the problem of out-of-date company classifications. By creating a vector-based classification (i.e., multi-label classification) we introduce the concept of similarity between companies as opposed to simplistic full assignment into single industry classes and sub-classes. We will see shortly in the next chapter the implications and use cases of this improvement.

Beyond designing an improved classification scheme we are faced by the challenge of implementing such a scheme. Even if theoretically superior a classification scheme is of little value without some means of implementation. Seemingly there is a trade-off between the complexity (or granularity) and the efficacy of a classification scheme. On one hand how well it represents the real world (with more classes, dimensions, highly structured and partial assigned) and on the other how useful and intuitive the scheme (with fewer labels, dimensions, etc). Whilst seemingly there are further extensions (multi-dimensional, hierarchical classification) we should consider how to best validate the efficacy of any classification scheme. In the abstract sense a classification scheme has no real value and the value they provide is derived from a particular use case or application.

In order to further assess the generated classification schemes (i.e., I_{Auto}, $I_{Multiple}$, $I_{Partial}$) we perform a user study with our industry partner Correlation Ventures, a venture capital firm (see Chapters 6 and 7). The next Chapter will utilise both existing and generated classification schemes in relation to use cases in the domain of venture finance, thereby, assessing their value and efficacy.

We are particularly interested in the investor-centric use cases of i) Identifying peers and ii) Matching investors $\leftrightarrow$ companies. Both the concepts of recommendation (i.e., Matching investors $\leftrightarrow$ companies) and similarity (i.e., Identifying peers) are highly relevant to the domain to recommender systems

and information retrieval. In our venture finance scenario we define *users* as venture capital firms (or individual investment partners) and *items* as investee private companies. Therefore our recommendation goals can be defined as i) defining company similarity measures (i.e., Identifying peers) and ii) the task of Top-N investment opportunity recommendation (i.e., Matching investors $\leftrightarrow$ companies). Whilst there have been some application of recommender systems to the broader domain of finance, including micro-finance [BS11], there has seemingly been no previous academic research in applying such techniques to venture finance.

We have chosen to focus on two pertinent use cases in the domain of venture finance:

- **Estimating Private Company Performance** — in order to analyse competitors and compare potential investment opportunities (see Chapter 6).
- **Top-N Investment Opportunity Recommendation** — for the purpose of finding co-investment partners and identifying the most suitable investors for a company and vice versa based upon past investment history (see Chapter 7).

It is important to note that whilst we are interested in these use cases for their own merit we are also intending to validate (or invalidate) the novel industry classification schemes generated.

Chapter 6

Estimating Private Company Performance

This chapter focuses on peer identification and estimating private company performance by observing the relationship between potential indicators and actual private company performance. Implementing a benchmarking methodology which allows for the comparison a company's performance against a set of relevant peers. Our methodology utilises the previously generated industry classification schemes, in order to improve private company similarity measures. Furthermore, we validate their real-world applicability in the domain of venture finance.

The funding environment for early-stage companies is ever changing. The dramatic decline in costs associated with starting a business gives rise to an increase in the number of prospective investment opportunity for a venture capital firm. Not solely the cost of operations but the means by which those companies seek and obtain capital to expand and grow is fundamentally changing.

6.1 Estimating Private Company Performance

It is against this backdrop that this section intends to investigate the scope for benchmarking peer-group performance of private companies. There is still great difficultly in assessing private companies performance, particularly in absolute terms, when many companies have little or no visible financial track record. Therefore the potential to make judgements based upon relative performance against a relevant peer-group using non-financial metrics becomes an attractive alternative.

Given the uncertainty of outcomes, lack of reliable performance data and cost of undertaking extensive due diligence it is difficult to analyse early-stage companies in isolation or in absolute terms. We therefore intend to focus on the relative performance of companies within comparable peer-groups over time and propose that there is an impetus for investors to specialise within a particular industry or sub-industry. Both of these broader aspirations are reliant on how we define the activities and relationships of companies, therefore improving our understanding of company classification is imperative.

In order to identify pertinent indicators of private company performance it is necessary to first

define what we mean by *performance* and thereby assess what may constitute a potential indicator of such performance.

Defining Performance

"Performance" along with notions of success and failure are subjective terms that demand context. They can be defined in absolute terms, for example successful may be defined as:

- *Company A is profitable*
- *Company A employs x number of employees*
- *Company A is acquired*

It is also possible to use relative measures, for example:

- *Company A has greater profits then Company B*
- *Company A employs more employees than Company B*
- *Company A was acquired for more than Company B*

Depending on the stakeholder in question there can be many different definitions of performance as shown in the Table 6.1:

Table 6.1: Definitions of Performance for Different Stakeholders.

Stakeholder	Motivation	Metric	Definition
Government and policy makers	Employment Job creation	High growth firms	A **high-growth firm** is firm with a minimum of ten employees at the beginning of a three-year period that achieves an average annualised employment growth greater than 20 per cent over that period.
Limited partners and institutional investors	Financial return	Net return	**Net return**, in relation to an investment manager, is the periodic return on assets under management, minus management fees and carried interest.
Fund managers	Financial return	IRR Multiples	**Internal Rate of Return (IRR)** is the discount rate at which the NPV of all cash flows related to an investment are equal to zero. **Multiples**, used in regard to a specific investment, means the total money returned divided by the total money invested (i.e., "for every dollar we invested, we got back X").
Portfolio company	Financial return Make an impact	Revenue Valuation[1]	**Revenue** is the money a company receives through its business activities, usually sales of a product or service. Common industry standard **Valuation** techniques include Current Value Method (CVM), Probability Weighted Expected Return Method (PWERM) and Option Pricing Method (OPM)[2].

Performance is often discussed in the context of growth. At a national level the indicator of a nation's economic health is growth in gross domestic product (GDP). In public markets an analyst may look for growth in earnings or in a company's stock price. Whilst the Organisation for Economic Co-operation and Development (OECD) defines a high-growth firm as "any firm with a minimum of ten employees at the beginning of a three-year period that achieves an average annualised employment growth greater than 20 per cent over that period", this measure is more relevant to governments and policy makers.

As previously stated, we are focused on the investors' perspective. If we again consider the role of venture capital firms seeking new ventures with high potential returns [Met07]. With this in mind we can look at the individual portfolio company, after all the potential return to an investor is derived from the underlying company value. Typically company value is measured either by using valuation techniques, which can be split into three main approaches: asset-based, income-based and market-based. Therefore a company's valuation is either based on its assets, forecasted cash flows, comparables or market transactions (i.e., investment round, acquisition, public offering). Unfortunately, in practice, reliable historical data on company valuation is hard to find and further confounded by the numerous different approaches to company valuation. There is also evidence to suggest that a multitude of other human factors effect valuation (e.g., economic actors, adverse selection, moral hazard) whereas revenue is arguably a purer measure of company performance.

An alternative approach would be to model a proxy such as revenue or even some success criteria (i.e., defined by a successful financial exit event of acquisition, merger or public offering). Whilst each has their inherent weaknesses (e.g., a company can easily grow revenues whilst never turning a financial profit) they benefit from having available historical data. There is also the question or whether we should be looking at absolute financial performance (e.g., Annual revenue) or other measures such as ratios and rates (e.g., Revenue per employee, Growth in annual revenue).

Potential Performance Indicators

The next important consideration is the set of potential input variables, which are to be modelled against either revenue or success. One possibility is reviewing academic literature on venture capital investment criteria (see Appendix D) as a source of additional pertinent dimensions for any proposed classification scheme. There are numerous different academic studies on investor criteria [TB84, Mul03, MHSM02, TS06, ZM98, CKL05a]. As noted [RW02] "the relative importance of selection criteria may vary between VCs and across different stages of the investment process," however, "most studies have reached similar conclusions regarding the relative importance of the various decision criteria used by VCs". With the most important being the managerial team, then market and product characteristics followed by expected financial outcomes [Man02].

Using the previously discussed literature on investment criteria (see Appendix D) it is possible to extract key metrics and assess the potential sources of relevant data for each. Table 6.2 outlines some of the proposed metrics and definitions:

A key question is figuring out which of these metrics are valuable, providing an indication of future revenue growth, and non obvious, therefore making their inclusion meaningful. Furthermore, any potential indicators need to be observable in order to be useful. Arguably, several of these metrics could also be used as different performance measures in their own right (e.g., Employee count, Active users, etc)

Table 6.2: Potential Indicators for Modelling Company Performance.

Criterion	Metric	Definition
Start-up team	Previous success	Successful Exit or IPO from previous company?
	Startup experience	Previous startup experience?
	Industry experience	Previous industry experience?
	% Ownership	%Share ownership of management team
	Employees	Number of full-time employees
Product service offering	IPR	Registered IPR?
	Visits	Website traffic in terms of unique visitors
	Users	Registered users
		Active users
Market/Industry Size	Growth	Total revenue of existing firms in market
	Market share	% Compound annual growth rate of market size
		Relative search queries compared to competitors?
Expected financial return	Valuation	Valuation
	Funding history	Previous investors?

as opposed to using a financial metric such as revenue.

6.1.1 Preliminary Experiments

Our aim was to model company performance in terms of revenue by using various web-based indicators, also referred to as a company's "digital footprint", as our inputs. Essentially we want to estimate company revenue based on their "digital footprint".

Data availability

A related issue is the availability of such data on private companies. Both in terms of actual performance and and potential indicators. An obvious source of performance data for UK companies is Companies House, the registrar of UK companies. All forms of companies are incorporated and registered with Companies House and file annual financial statements in addition to annual company returns, which are all public records under the relevant law[3]. However, Companies House maintains certain thresholds for which companies have to file annual statements and returns:

- Exemption
  - have operated for less than 21 months
- Abbreviated accounts
  - have a turnover of not more than £6.5 million; and
  - have a balance sheet total of not more than £3.26 million

Therefore little or no revenue data will be accessible for companies which have operated for less than 21 months or fall under the above financial threshold.

Table 6.3 depicts the broad types of data sources, examples of specific sources and other informations about the coverage, time period and potential issues:

In terms of sources for potential performance indicators the scope is much broader, however, as displayed in Table 6.3 below, the various sources offer limited access to historical data.

[3] Companies Act 2006 — http://www.legislation.gov.uk/ukpga/2006/46/contents

Table 6.3: Data Sources and Potential Issues.

Type	Example	Time period	Potential issues
Social media data	Twitter (Follower counts) LinkedIn (Employee mentions)	<1 year	Limited historical access (e.g., Twitter previous 60 days, Klout previous 30 days, LinkedIn no historical data)
Accounting data	Companies House (Assets, Liabilities, Employees)	<5 years	Different accounting practices Small company exemptions
Web traffic / Search data	Alexa Compete Google Trends	<10 years	Limited historical access (e.g., Google Trends since 2004, Alexa since 2007) Sample sizes used in web traffic estimate methodologies
Venture financing data	CrunchBase (Funding rounds, Exits) VentureSource	<20 years	Survivor bias Incomplete (i.e., not all firms are venture backed)

"Digital footprint"

Given the difficultly in finding data on private companies that is both accessible (i.e., easy to harvest) and has broad coverage (i.e., covers a large number of private companies) a brief partnership with a London-based company called Startup Intelligence (registered as Pelucid Limited) offered access to such data. The Startup Intelligence dataset offered two main types of data:

- Financial data — data from Companies House annual filings include in financial statements and annual returns (e.g, Revenue, Total Assets, Total Liabilities)

- "Digital footprint" data — a selection of web-based data related to companies online presence (e.g, Alexa Rank, SEMrush Organic Cost, Facebook Shares)

Clearly this offered a potential performance measure (i.e., Revenue) and a host of potential performance indicators to be modelled. Most of these data points served as proxies or surrogate measures for underlying metrics we are interested. For example AlexaRank is an estimation of a website's traffic or number of visitors. Another important note is that most of the companies in the dataset related to a particular office space operated by a company called the Workspace Group. This presented a potential issue of sample bias to the type of companies (i.e., largely small and medium enterprises) included in the dataset.

Experimental Setting

There are broadly two approaches to achieving this with both classification (i.e., discrete) and regression (i.e, continuous).

Estimating company revenue from their "digital footprint" using classification

Using annual revenue as our measure of performance, we attempt to model company performance discretisation using various potential indicators. Our aim is classifying companies into bins based upon annual revenue using their "digital footprint" as our input. Abstractly, the classification model (e.g., Decision Tree) is learned from a training dataset by deciding which attribute is the most helpful for classifying the different companies (i.e., information gain). In order to achieve this we used a sample of around 200 London-based companies each with 25 input attributes and a value for annual revenue (i.e.,

target attribute or label). The main algorithms used were decision tree (c4.5) and random forest (W-RandomForest) which had the benefits of being "white box" algorithms (c4.5 only) and also the capacity to handle categorical input features (e.g., industry) without dummy encoding (see Appendices).

Estimating company revenue from their "digital footprint" using regression

The process for estimating company revenue using regression (i.e., to predict a continuous output) was much the same as the workflow for classification and therefore will not be repeated in detail. The main differences we the following:

- No discretisation — our target variable revenue is no longer discretised into separate bins or ranges of revenue, instead the continuous value (e.g., £2,500,000) is used
- Techniques used — instead of decision tree (c4.5) and other classifier algorithms the following regression techniques were used:
  - Linear Regression — calculates a linear regression model using the Akaike criterion for model selection
  - LibSVM — libSVM implementation of a support vector machine learner for classification
  - k-NN — k—nearest neighbour implementation based on an explicit similarity measure

Evaluation Measures

In order to evaluate the accuracy of the classification we perform a ten-fold cross validation (X-validation) and report average % performance metrics. The input example set is split up into number of validations subsets using stratified sampling which builds random subsets and ensures that the class distribution in the subsets is the same as in the whole example set. These subsets are then used to train and test our model by applying the trained model and measuring its performance (i.e., Accuracy). The same cross validation is used for regression and instead of reporting accuracy (as with classification) we had an error term in the form of mean-squared error.

Preliminary Results

Table 6.4 provides an overview to the initial results from testing different classifiers with the aim of auto-classifying companies into discrete revenue ranges based upon the company's "digital footprint".

Table 6.4: Preliminary Classification Results.

Technique	Examples	Attributes	Accuracy %	Notes
Decision Tree (c4.5)	195	24 regular, 1 label	8.71%	Signals only
Weka Random Forest	195	24 regular, 1 label	15.42%	Signals only
Decision Tree (c4.5)	195	29 regular, 1 label	20.45%	Signals and SI classification
Weka Random Forest	195	29 regular, 1 label	47.68%	Signals and SI classification

The main observations were as follows:

- Random Forest achieved increased accuracy over Decision Tree (c4.5) classifier

- Cross-validation shows poor accuracy ($\sim 10\%$) in classifying companies into revenue ranges
- This accuracy was improved by using a peer-group classification alongside the "digital footprint" data

Figure 6.1 illustrates predicted annual revenue (y-axis) versus actual annual revenue (x-axis) (with colour denoting company age in months) shown on a log scale. By inspection we can see that estimating company revenue for companies with annual revenue less than £10,000,000 is problematic and largely inaccurate.

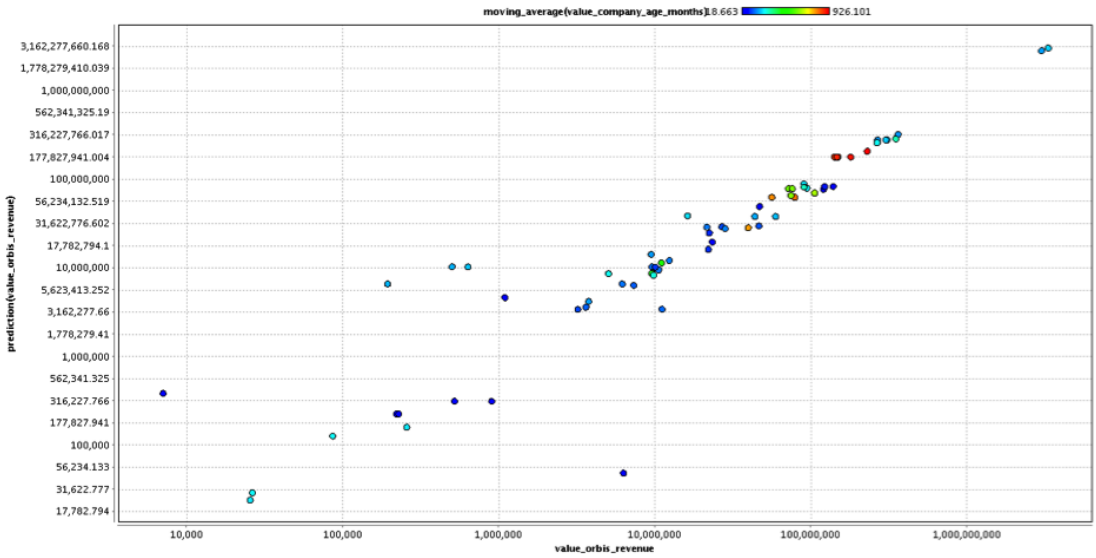


Figure 6.1: Plot of Predicted and Actual Annual Revenue.

Table 6.5 below provides an overview to the initial results from testing different regression technique with the aim of predicting companies annual revenue.

Table 6.5: Preliminary Regression Results.

Technique	Examples	Attributes	Mean squared error
Linear Regression	227	95 regular, 1 label	3732225630.890
LibSVM	227	95 regular, 1 label	988723638.253
k-NN	227	95 regular, 1 label	190297590.535

Using the "digital footprint" data a rudimentary sensitivity analysis was performed. The attributes were grouped in their sources (e.g., Alexa) and then the original classification approach to revenue estimation was executed with each attribute both independently and with all attribute except that particular group. Table 6.6 shows the accuracy of classification when used independently and effect of removing an attribute group.

The outcome of this analysis is inconclusive but it gives some indication of the significance or predictive power of each attribute group. A more thorough attribute-by-attribute sensitivity analysis may prove more useful. Despite these findings it is difficult to say what accuracy is "good enough" to have value to investors?

Table 6.6: Sensitivity Analysis of Attribute Groups.

Attribute group	Accuracy used independently	Effect of removing feature group
Alexa	66.88%	+ 3.55%
Bing	59.02%	+ 1.55%
Facebook	54.37%	+ 0.8%
Indeed	27.57%	- 0.37%
LinkedIn	71.18%	- 0.03%
SEMrush	63.37%	+ 0.77%

Issues

In general, further development and validation of techniques for modelling performance was deemed necessary but we also needed to address several challenges. Particularly missing data (i.e., >50% missing data for many attribute groups) and non-contiguous time period (i.e., a longitudinal sample (i.e., company history over time) with signals overlapping financial data is currently not available).

Discrepancy amongst time intervals (i.e., Daily, weekly, monthly, yearly), types (i.e., Discrete-continuous; Nominal-numeric) and measurement (i.e., Cumulative, non-cumulative) was resolved by attempting to standardise signals by creating single value per signal (e.g. daily average over prior 12 months) and transforming nominal to numeric data if necessary (e.g., LinkedIn industry/sector attribute for use with SVM).

Whilst some of the above issues would be easily remedied by further preprocessing and smoothing the issue of non-overlapping data, essentially non-contiguous time-series, is seemingly unresolved given the current dataset. The solution proposed of using companies with revenue filed in past 12 months only makes a large assumption that companies revenue shows little variance over a 12-month period. However, given we have access to past revenue data we know the assumption to be false.

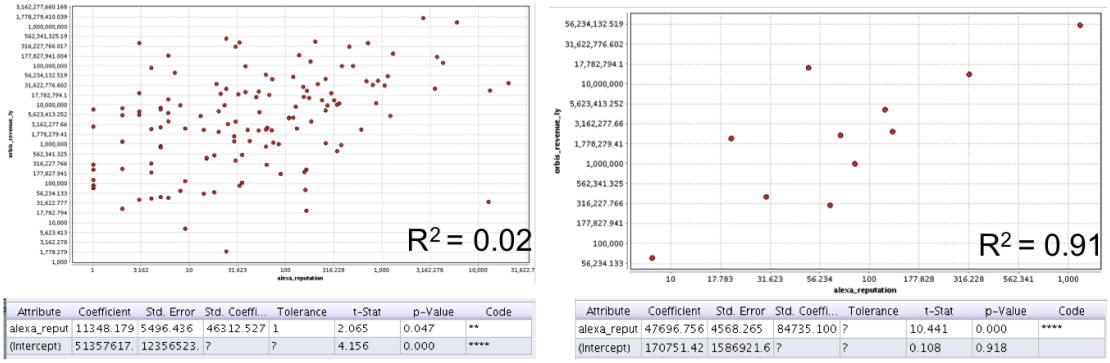
6.2 Peer Groups

When analysing early-stage private companies, in the absence of rich performance data, many look at comparables. It is difficult to analyse a company in isolation on absolute terms, therefore it is logical to attempt to identify peer-groups of similar companies and observe relative measures of performance. The term “peer-group” originates in the late 1940s and is more commonly used in relation to child development literature, however, it gives a good comprehension of how we envision company classification.

Even in public markets where investors are overwhelmed by rich data on both fundamental (e.g., quarterly financial statements) and technical (e.g., market prices and trading volumes) aspects of publicly traded company performance many use comparative measures such as price-earning (PE) ratios.

A simple illustration of the motivation for using peer-groups can be provided by looking at a number of companies and plotting some performance measure (e.g., revenue) against some potential performance indicator (e.g., number of inbound links to company’s website). Observing a relatively small population of companies (n =205) there is seemingly little or no relationship between the two variables. However if we define a peer-group (e.g., Hotels & Accommodation) there is some indication of a linear correlation

between the two variables.



Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
alexa_reput	11348.179	5496.436	46312.527	1	2.065	0.047	**
(Intercept)	51357617.	12356523.	?	?	4.156	0.000	***

Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
alexa_reput	47696.756	4568.265	84735.100	?	10.441	0.000	***
(Intercept)	170751.42	1586921.6	?	?	0.108	0.918	

Figure 6.2: Illustration of Peer-group using Linear Regression.

Whilst, with such a small sample peer-group (n=11) any correlation is not statistically significant, it gives some illustration of the intention of using peer-groups in order to compare similar companies against relevant metrics.

In the context of an investor considering an investment opportunity we would likely want to consider the competitive landscape and especially any direct competitors or "peers" for a particular company. Ideally we would want to identify similar companies in terms of their business models, customers and market segments, life stage, location and other pertinent factors. However, clearly there are limitations imposed by available datasets and information on business models and life stage may not always be readily accessible. Therefore, our focus is on textual descriptions of companies activities and their industry classification.

6.2.1 Datasets

CrunchBase Dataset

CrunchBase describes itself as a "free database of technology companies, people, and investors that anyone can edit" [Cru13a]. The CrunchBase dataset includes a generated principal component representation of the investee private companies' descriptions and an industry category code used by CrunchBase to depict industry sectors (e.g., "Games, Video & Entertainment", "Mobile"). Data was access via publicly available API[4] (Application programming interface). Table 6.7 depicts the relevant data points for each entity (i.e., company).

Table 6.7: CrunchBase Dataset.

Attribute	Description
entity_id	Entity (i.e. company) identity
prin[1-10]	Principal components derived from a singular value decomposition (SVD) of CrunchBase overview
name	Entity (i.e., company) name
overview	Entity CrunchBase overview (i.e., textual description)
entity_permalink	Entity CrunchBase permanent link
category	CrunchBase category classification

[4] CrunchBase Developer Portal —http://developer.crunchbase.com

VentureSource Dataset

With VentureSource in total we have 21,610 investee private companies with a generated principal component representation of the investee private companies' descriptions and three levels of hierarchy for industry classification: industry groups (1st level); industry segments (2nd level); and industry codes (3rd level). Our dataset was provided by Dow Jones VentureSource, a leading data provider to the venture capital industry, courtesy of Correlation Ventures.

Table 6.8: VentureSource Dataset.

Attribute	Description
e_entityid	Encrypted entity (i.e., company) identity
prin[1-10]	Principal components derived from a singular value decomposition (SVD) of VentureSource company description
new_industrygroup	VentureSource industry group (1st tier) classification
new_industrysegment	VentureSource industry segment (2nd tier) classification
new_industrycode	VentureSource industry code (3rd tier) classification

Whilst VentureSource is our primary dataset we also have access to a secondary dataset from CrunchBase. VentureSource is anonymised due to privacy restrictions, therefore, CrunchBase is often used for illustrative purposes. Detailed results for both datasets are included in the Appendix C. As shown in Table 6.8 it is important to note that the VentureSource classification scheme has a hierarchical structure with three tiers (Group, Segment and Code see 5.2.1 for more detail). Group describes broad industry sectors (e.g., "Information Technology") and the subsequent tiers Segment (e.g., "Software") and Code (e.g., "Communications Software") provide further granularity (see Section 5.2.1 for more detail).

6.2.2 Existing and Novel Industry Classifications

Table 6.9: Summary of Existing and Generated Classification Schemes.

Classification Scheme	$I_{Existing}$	I_{Auto} (Auto)	$I_{Multiple}$ (Multi)	$I_{Partial}$ (Partial)
VentureSource	vs_group vs_segment vs_code	vs_group_auto vs_segment_auto vs_code_auto	vs_group_multi vs_segment_multi vs_code_multi	vs_group_partial vs_segment_partial vs_code_partial
CrunchBase	cb_category	cb_category_auto	cb_category_multi	cb_category_partial
AngelList	al_tag	al_tag_auto		al_tag_partial

In the previous chapter we described a number of both existing and generated (i.e., auto-classified, multi-dimensional) classification schemes to be used as inputs for our evaluation against our intended evaluation use-cases, a summary of the various schemes is outlined in Table 6.9. We utilise both existing (e.g., Group, Segment, Code, Full[5]) and generated industry classification vectors (i.e., Auto, Multi, Partial) as item category information and also their description in the form of principal components (i.e., PC) to base our item-item similarity measures for our recommendation models (see Section 7.3.2 and 7.3.3).

Furthermore, we apply collaborative filtering (CF) by observing the overlap of the VC firms (or

[5] Full refers to using the entire VentureSource hierarchy as a single vector rather than a specific tier.

investment partners) of investee companies as an alternative item-item similarity measure. The logic being that private companies with similar investors (i.e., VC firms or investment partners) are similar. This is reliant on the notion that investors specialise to some extent which is discussed in detail later (see Section 7.2).

6.3 Experiments

We have designed a user study in collaboration with industry partner Correlation Ventures, a venture capital firm, in order to test our proposed use case in venture finance:

- **Peer Identification and Company Similarity Measures** — in order to analyse competitors and compare potential investment opportunities.

6.3.1 Experimental Settings

As a preliminary validation of the various existing and generated classification schemes we have devised a user study with a small number of investment partners at Correlation Ventures.

We make use of our access to the VentureSource dataset and test two different company similarity measures and benchmark against selecting a random subset of companies with the same industry Code (i.e., the 3rd tier) of the VentureSource industry classification scheme. Therefore our compared techniques for identifying the top 8 most similar companies or "peers" are:

1. Random subset with the same industry Code as exemplar company
2. Using cosine similarity over principal components PC vector (i.e., no classification scheme)
3. Using cosine similarity over $I_{Partial}$ industry vector

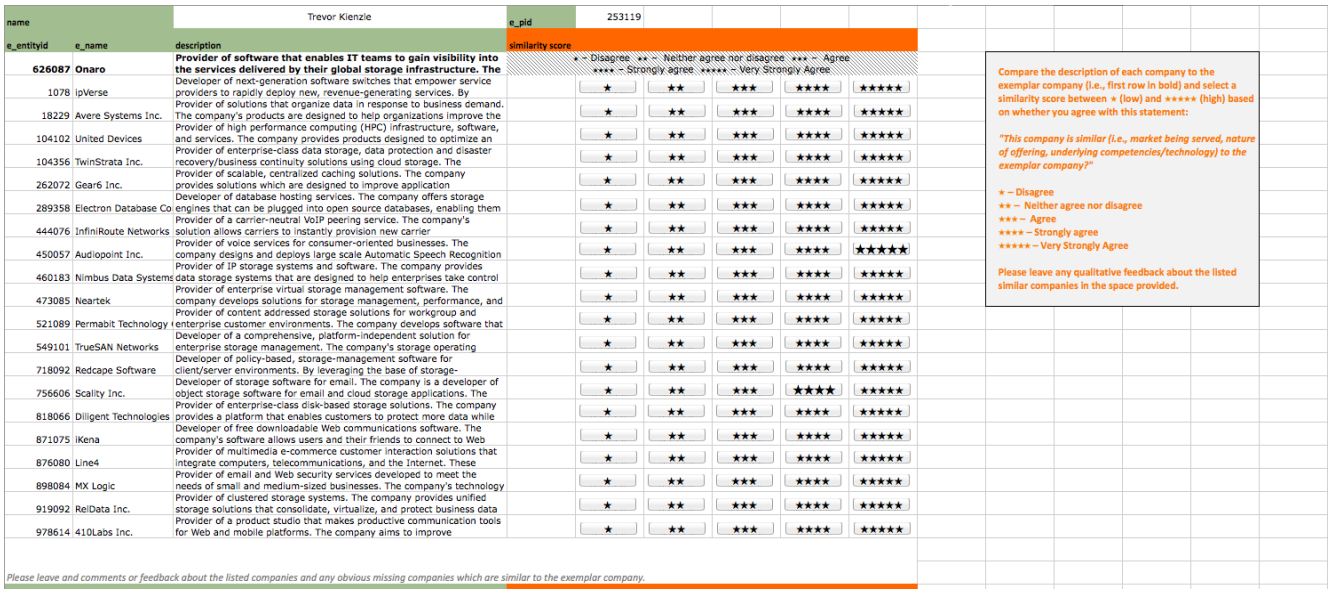
We then selected 8 exemplars companies from VentureSource and using the above techniques we compute similarity measures and select the 8 most similar companies for each technique.

6.3.2 Experimental Measures

Each investment partner is then provided as worksheet with the exemplar companies, peer groups and a field for providing a similarity score between 1 (low) and 5 (high). Respondents are required to compare the description of each company to the exemplar company and select a similarity score between $\star$ (low) and $\star\star\star\star\star$ (high) based on to what extent they agree with the following statement: *"This company is similar (i.e., market being served, nature of offering, underlying competencies/technology) to the exemplar company"*. Finally, respondents are requested to leave any qualitative feedback about the listed similar companies. An example of the user study worksheet is shown in Figure 6.3.

6.3.3 Experimental Results

The outcome of the initial user study into Peer Identification and Company Similarity Measures:



name		Trevor Kienzle	e_pid	253119				
e_entityid	e_name	description	similarity score					
626087	**Onaro**	**Provider of software that enables IT teams to gain visibility into the services delivered by their global storage infrastructure. The**	★ – Disagree ★★ – Neither agree nor disagree ★★★ – Agree ★★★★ – Strongly agree ★★★★★ – Very Strongly Agree					
1078	ipVerse	Developer of next-generation software switches that empower service providers to rapidly deploy new, revenue-generating services. By	★	★★	★★★	★★★★	★★★★★	
18229	Avere Systems Inc.	Provider of solutions that organize data in response to business demand. The company's products are designed to help organizations improve the	★	★★	★★★	★★★★	★★★★★	
104102	United Devices	Provider of high performance computing (HPC) infrastructure, software, and services. The company provides products designed to optimize an	★	★★	★★★	★★★★	★★★★★	
104356	TwinStrata Inc.	Provider of enterprise-class data storage, data protection and disaster recovery/business continuity solutions using cloud storage. The	★	★★	★★★	★★★★	★★★★★	
262072	Gear6 Inc.	Provider of scalable, centralized caching solutions. The company provides solutions which are designed to improve application	★	★★	★★★	★★★★	★★★★★	
289358	Electron Database Co	Developer of database hosting services. The company offers storage engines that can be plugged into open source databases, enabling them	★	★★	★★★	★★★★	★★★★★	
444076	InfiniRoute Networks	Provider of a carrier-neutral VoIP peering service. The company's solution allows carriers to instantly provision new carrier	★	★★	★★★	★★★★	★★★★★	
450057	Audiopoint Inc.	Provider of voice services for consumer-oriented businesses. The company designs and deploys large scale Automatic Speech Recognition	★	★★	★★★	★★★★	★★★★★	
460183	Nimbus Data Systems	Provider of IP storage systems and software. The company provides data storage systems that are designed to help enterprises take control	★	★★	★★★	★★★★	★★★★★	
473085	Neartek	Provider of enterprise virtual storage management software. The company develops solutions for storage management, performance, and	★	★★	★★★	★★★★	★★★★★	
521089	Permabit Technology	Provider of content addressed storage solutions for workgroup and enterprise customer environments. The company develops software that	★	★★	★★★	★★★★	★★★★★	
549101	TrueSAN Networks	Developer of a comprehensive, platform-independent solution for enterprise storage management. The company's storage operating	★	★★	★★★	★★★★	★★★★★	
718092	Redcape Software	Developer of policy-based, storage-management software for client/server environments. By leveraging the base of storage-	★	★★	★★★	★★★★	★★★★★	
756606	Scality Inc.	Developer of storage software for email. The company is a developer of object storage software for email and cloud storage applications. The	★	★★	★★★	★★★★	★★★★★	
818066	Diligent Technologies	Provider of enterprise-class disk-based storage solutions. The company provides a platform that enables customers to protect more data while	★	★★	★★★	★★★★	★★★★★	
871075	iKena	Developer of free downloadable Web communications software. The company's software allows users and their friends to connect to Web	★	★★	★★★	★★★★	★★★★★	
876080	Line4	Provider of multimedia e-commerce customer interaction solutions that integrate computers, telecommunications, and the Internet. These	★	★★	★★★	★★★★	★★★★★	
898084	MX Logic	Provider of email and Web security services developed to meet the needs of small and medium-sized businesses. The company's technology	★	★★	★★★	★★★★	★★★★★	
919092	RelData Inc.	Provider of clustered storage systems. The company provides unified storage solutions that consolidate, virtualize, and protect business data	★	★★	★★★	★★★★	★★★★★	
978614	410Labs Inc.	Provider of a product studio that makes productive communication tools for Web and mobile platforms. The company aims to improve	★	★★	★★★	★★★★	★★★★★	

Please leave and comments or feedback about the listed companies and any obvious missing companies which are similar to the exemplar company.

Compare the description of each company to the exemplar company (i.e., first row in bold) and select a similarity score between ★ (low) and ★★★★★ (high) based on whether you agree with this statement:

"This company is similar (i.e., market being served, nature of offering, underlying competencies/technology) to the exemplar company?"

★ – Disagree
★★ – Neither agree nor disagree
★★★ – Agree
★★★★ – Strongly agree
★★★★★ – Very Strongly Agree

Please leave any qualitative feedback about the listed similar companies in the space provided.

Figure 6.3: User Study Worksheet for Peer Identification.

- On average both techniques 2 (i.e., Cosine similarity over PC vector) and 3 (i.e., Cosine similarity over $I_{Partial}$ vector) improve over technique 1 (i.e., Random with same Code)
- Similarity scores very close between techniques 2 and 3
- Both 2 and 3 perform poorly with limited or brief exemplar descriptions (e.g., "Management company for The Foundry LLC, a medical-device incubator.")

The results of the two individual investment partners in the study were directionally very similar but with lower scoring across the board by one investment partner. Overall, technique 2 (i.e., Cosine similarity over PC vector) and 3 (i.e., Cosine similarity over $I_{Partial}$ vector) seemingly exceed technique 1 (i.e., Random with same Code) when there is a reasonable textual description. However, there seemingly not a great difference between using the partially assigned industry representation versus the principal component representation. Clearly, further user studies beyond this pilot with a larger number of participants are necessary to validate these initial findings.

6.4 Discussion

Whilst we have initiated research into defining and estimating private company performance based on a "digital footprint" (i.e., publicly available company data) working alongside an industry partner Startup Intelligence. Unfortunately due to limitations largely in relation to access and availability of longitudinal datasets this area of research remains incomplete. We have identified use cases for such an undertaking including screening prospective opportunities and benchmarking peer-group performance either pre- or post-investment. We have also discussed the importance of several key considerations including

company life stage, business model and whether there are good proxy measures or indicators for all important investment criteria (e.g., Management team, market opportunity).

6.4.1 Benchmarking Performance

Our goal in benchmarking private companies against a relevant peer-group (i.e., group of similar companies) is to better understand their relative performance. Whether a company is over- or under- performing relative to its peers. To date two different potential methods of benchmarking have been considered. The first is comparing companies using a weighted composite (or unified) score. The second is to compare companies against individual metrics which are deemed important to that specific peer-group. The motivation is to attempt to quantify potential indicators of company performance relative to a peer-group of similar companies.

Using A Weighted Score

Essentially VentureRank is simply a weighted score in order to compare a company against a peer group:

$$VentureRank = w_1\left(\frac{x_1}{x_{1_{max}}}\right) + w_2\left(\frac{x_2}{x_{2_{max}}}\right) + \cdots + w_n\left(\frac{x_n}{x_{n_{max}}}\right)$$

Figure 6.4: VentureRank Weighted Scoring Formula.

In a simplistic scenario weights may be defined using linear regression:

$$Y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \epsilon \tag{6.1}$$

Where Y represents the independent variable annual revenue and the following example dependent variables x_1 Alexa reputation (i.e., number of inbound website links), x_2 Indeed jobcount (i.e., number of online job posts), x_3 LinkedIn number followers (i.e., number of company followers on LinkedIn) and x_4 Facebook shares (i.e., number of times website shared on Facebook). Using linear regression, for example (or other approaches), we can define the relative weights, w, of each input variable in our "digital footprint" data.

Using a worked example, if we define our peer-group as follows:

- Sales model = Subscription
- Product type = Web application
- Sector context = IT
- Client type = Businesses
- Retail premises = No

We can then generate a weighted VentureRank score for each company:

Table 6.10: Worked Example of VentureRank.

Company ID	x_1, Alexa reputation	x_2, Indeed jobcount	x_3, LinkedIn number followers	x_4, Facebook shares	VentureRank
1127 94	0	38	14	0	.10
4667	225	2	142	2	.30
4851	1449	0	239	74	.90
6383	462	1	124	14	.35
w	0.5	0.1	0.2	0.2	1

In this example we have only four companies each with values for the various input variables $x_1 \ldots x_4$ and we have assigned weights, w to each variable respectively to then calculate a VentureRank score for each company which is a value between 0 and 1, alternatively it can easily be scaled to between 0 and 100 for easy interpretation. The higher the VentureRank score the better the company performance based upon the "digital footprint" data as a proxy for their actual financial performance (i.e., Revenue). It is important to consider that unifying performance into a single metric such as VentureRank may not be desirable. Some studies on venture capital selection criteria find that typically VCs screening criteria are conjunctive and non-compensatory [RR92]. Anecdotal evidence suggest VCs prefer to invest on the basis of excellence in a key dimension (e.g., team, technology, traction) as opposed to companies which are merely satisfactory or even above average across several dimensions. Most likely due to the extremely skewed returns distribution and the "winner takes all" mentality of venture capital investments.

Using Peer-group Specific Metrics

Perhaps more important is defining appropriate peer-group and then comparing a single meaningful metric for that peer-group. For, example a commonly applied metric in relation to mobile apps is "stickiness" defined as the ratio of daily active users (DAU) to monthly active users (MAU). Some overarching metrics are most likely relevant to all companies (e.g., Revenue) but others are unique to the business model and therefore peer-group under assessment, for example subscription-based models (e.g., Average Revenue per User (ARPU)). The metrics to be defined and monitored should be the ones that related to a particular business model [CY13]. Therefore, it may be preferable as opposed to weighting various metrics across all different companies and peer-groups.

6.4.2 Company Life Stage

There is potential to further define or inspect companies by life stage dimension. This could defined by using thresholds in certain metrics but also by milestones. There are multiple key milestones in a company lifetime along different dimensions or aspects of the business, for example: Team (Founding team, Key executives); Product (Alpha, Beta, Launch); Market (First paying customer, Product-market fit); Expected financial return (Break-even, Cash-flow positive); Investment (Seed, Series A, Series B, etc); Exit event (Acquisition, Public offering).

Beyond explicit milestones there are also other different life-stage or life-cycle definitions from

management literature:

Table 6.11: Existing Definitions of Life-stage and Life-cycle.

Name	Stages	Source
Organizational Life Cycle	Birth Growth Maturity Decline Death	[Daf07]
Technology Adoption Lifecycle	Innovators Early adopters Early majority Late majority Laggards	[RB57, Moo99]
Marmer Stages	Discovery Validation Efficiency Scale Sustain Conservation	[MHD+12]

In this chapter we have illustrated a use case of our generated classification schemes from the previous chapter. Namely identifying peers in order to compare and benchmark their performance. Whilst our user study has to some degree validated the improved ability to identify peers using a vector-based classification schemes (e.g., $I_{Multiple}$, $I_{Partial}$) further user studies are necessary. We have also defined methodologies for benchmarking company performance and identified potential extensions and improvements for future research (i.e., Multi-dimensional classification, Company life stage).

The following chapter covers another use case for our generated classification schemes and the application of recommender system techniques in order to identifying relevant investment opportunities for venture capital investors based upon both industry focus and past investments.

Chapter 7

Top-N Investment Opportunity Recommendation

This chapter focuses on demonstrating the efficacy of recommender systems in relation to this novel application. Our methodology takes advantage of our access to venture financing data plus existing and generated industry classification schemes, ultimately seeking to improve the performance of Top-N investment opportunity recommendation.

As discussed previously, early-stage investment is characterised by: uncertainty of outcomes for early-stage companies; a lack of reliable data on private company performance; and associated cost of undertaking due diligence. Faced with a large number of prospective investment opportunities venture capital firms and their investment partners require some form of screening in order to undertake further due diligence and evaluation of the most attractive opportunities. Unfortunately, given the uncertainty and imperfect information this is a non trivial task. In such screening there is a danger of both false positives (i.e., a poorly performing investment) and false negatives (e.g., missing the next Google). This challenge has become increasingly difficult given the recent dramatic reduction in the costs of starting a company and the subsequent increase in new company formations.

Referral from trusted sources (e.g., entrepreneurs, accountants, lawyers, other investors) is often used to screen the seemingly infinite number of opportunities seeking further analysis and evaluation. However, with increasing globalisation not all promising investment opportunities are likely to be directly connected to any particular investor, however well networked they might be. Therefore, relying on referral from first or second degree connections is sub-optimal. Recommender systems provide a potential complement to traditional screening methods.

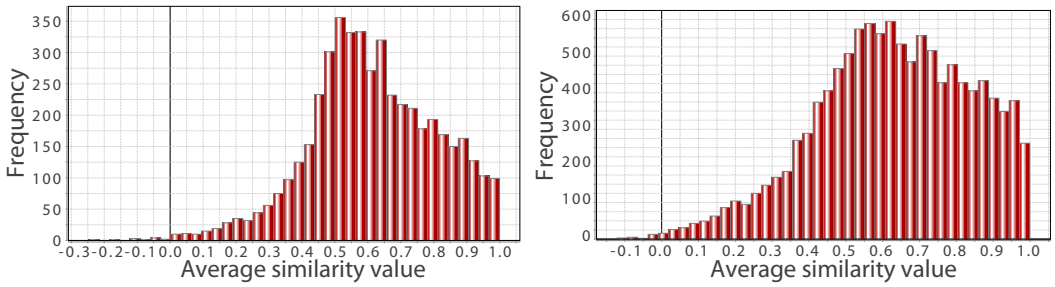
7.1 Recommender Systems

Recommender systems have emerged as an effective way to help people cope with the problem of information overload by providing personalised recommendations based on user and item profiles. In relation to venture finance the *user* being the investment firm (or individual investment partner) and the *item* being the private company.

Based on prior studies [RGFST11, Mon03, BFG11] it is possible to divide recommender system techniques into three main categories: content-based (or feature-based), collaborative-based (i.e., collaborative filtering) and hybrid approaches. In our experiments (see Section 7.4 we have compared content-based, collaborative-filtering (CF) and hybrid approaches. Content-based techniques make use of user or item attributes, in the context of venture finance this could be, for example, fund size (i.e., of the VC firm) or industry (i.e., of the private company). In contrast, collaborative-based techniques purely utilise interactions between users and items, in relation to venture finance this interaction investment. Further surveys on recommender systems focus on content-based [PB07] and collaborative filtering [SK09] respectively. We are particularly interested in hybrid approaches which both analyse interactions between the user [SK09] and items along with the item meta data (e.g., textual attributes, categorical attributes) or features [BYRN99]. Such hybrid approaches are well suited to our task of investment opportunity recommendation and our desire to utilise industry classification information.

In this context, we are particularly interested in content-based techniques [BYRN99] which analyse interactions between a particular user and a set of items using item meta data (e.g., textual attributes, categorical attributes) or features. Whilst content-based recommender systems are domain-dependent and do not tend to provide serendipitous recommendations for users [AfSG11] they do not suffer from issues faced by collaborative-based techniques (e.g., cold-start problem) [SPUP02]. The cold-start problem is well known in the domain of recommender systems research. Essentially, when a new user (or item) has no previous interactions there is no association between such a user (or item) and any other items (or users). In the context of venture finance, if a new venture capital fund is established initially prior to investing we have no track record of investments with which to infer relevant investments, at least using collaborative filtering techniques, hence the term “cold start”.

Given the sparsity of collaborative data for our particular use case with venture capital firms making only a small number of investments each year and with limited co-investment content-based (or hybrid approaches) recommender systems seem appropriate. Although we are interested in improving the accuracy and relevance of recommendations in our particular use cases (see Section 5.6.2) we are also interested in evaluating the utility of different classification schemes. The existing and generated industry classification schemes (see Chapter 5) cannot be evaluated in any abstract sense as “better” or “improved” unless in relation to some specific use case in venture finance. Therefore we intend



(a) VC firm (b) Investment partner

Figure 7.1: Average Pair-wise Cosine Similarity for VC and Investment Partner Portfolios.

to compare and contrast the different industry classification schemes against our particular use case of investment opportunity recommendation described in the following section.

7.2 Specialisation by Industry

We are interested in the decision making trade-offs made by investors (e.g., to specialise or diversify) under conditions of uncertainty. In particular discovering whether VC firms and their individual investment partners specialise in terms of industries or sub-industries in which they make their investments. Intuitively we would expect individual investment partners, and to a lesser extent VC firms, to specialise in their investment strategies.

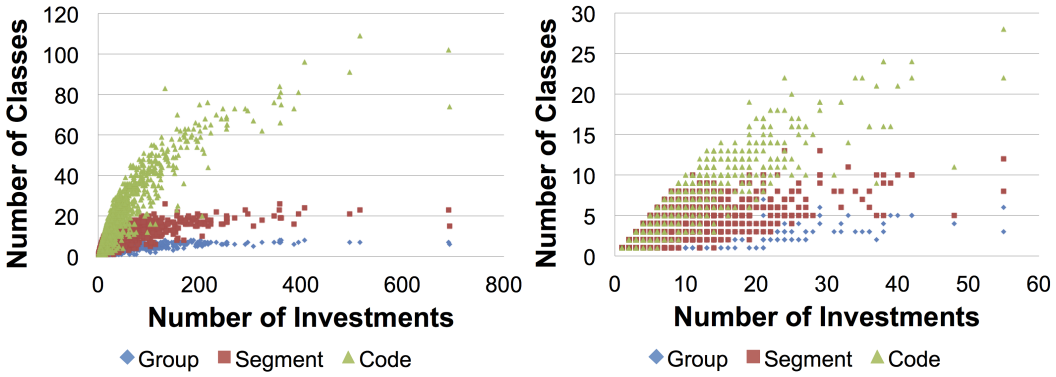
Whilst there are is no strict limits an "average" VC firm (i.e., $100 million fund size) will have a small number (i.e., less than 10) of investment partners who will take board seats in the companies in which they chose to make investments. These individual investments constitute the VC firm's overall portfolio of investments, which we would, again only intuitively, expect to be specialised to some degree, at least beyond a random sample of private companies. For a portfolio P of n companies we calculate $n(n-1)/2$ similarity measures aggregated using an average pair-wise similarity , s, across the portfolio, based upon the principal components derived from company descriptions.

Table 7.1 depicts a randomly selected portfolio alongside a randomly selected VC firm (vcid=221056) and investment partner (pid=733043) with their respective measures which agree with our hypothesis around specialisation. Furthermore, Figure 7.1 shows the distribution of average pair-wise cosine similarity s for portfolios across all VC firms and all investment partners. The positive skew of the distributions, suggest the dominant investment strategies are in favour of specialisation, especially for individual investment partners. However, some VCs and investment partners have negative measures suggesting specialisation although the norm is not always the chosen investment strategy.

An initial analysis was undertaken to observe the number and industry specialisation of investments by VC firms and investment partners. In Figure 7.2 investments are plotted against the count of unique industry classes they cover at the three different levels of the hierarchy. We see a concentration of investments in a small number of industry classes, particularly for investment partners, even at the lowest

Table 7.1: Specialisation of Example Portfolios.

Portfolio, P	Average pair-wise cosine similarity of principal components, s
Random portfolio	0.326 ± 0.241
vcid=221056	0.516 ± 0.230
pid=733043	0.758 ± 0.133



(a) VC firm (b) Investment partner

Figure 7.2: Number of Investments Against Number of Classes by User.

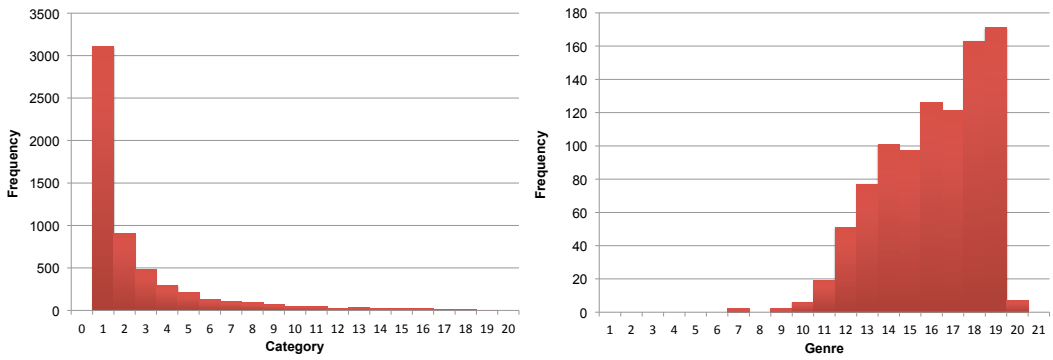
level of the industry hierarchy (i.e., Code).

Comparison To Other Domains

Many of the approaches (e.g., collaborative filtering, content-based) and techniques (e.g., neighbourhood methods, latent factor models) have been developed and applied rigorously in the context of movie recommendation using academic datasets such as MovieLens and Netflix. Given the richness of research in this domain it is useful to provide some comparison between the properties and characteristics of these two distinct domains through exploratory data analysis.

Comparing the distribution of unique categories between CrunchBase and MovieLens[1], we identify quite different characteristics. As shown in Figure 7.3, the investors in CrunchBase tend to interact with companies belonging to a small number of industry categories. On the contrary, the users in Movielens have probably watched a variety of movies, often across more than 15 different genres. Therefore, we can infer that for a movie recommendation task, the movie genre information may not be so effective since users always have an interest across diverse genres, whilst for the investment opportunity recommendation, the industry category information of companies will be more effective because investors tend to focus on fewer industry categories. Such differences motivate us to explore leveraging industry classification information (in Section 7.3).

[1] We choose Movielens here because it contains genre information of movies while Netflix does not.



(a) CrunchBase (b) MovieLens

Figure 7.3: Unique Category Distribution by User.

7.3 Top-N Recommendation

Given the increasing number of potential investment opportunities for consideration there is scope for better matching relevant opportunities to investors based upon their past investment activities. Whilst further in-depth evaluation is important the initial origination and screening of investment opportunity may be improved through the application of recommender systems.

7.3.1 Datasets

CrunchBase Dataset

In additional to the indsutry classification data (see Section 6.2.1) for both CrunchBase and VentureSource we have the additional data providing information about past investment history for VC firms and investment partners.

CrunchBase [Cru13a] includes historical venture financings in the United States. Again, for consistency, we focus on US-based companies, which have received investment in US dollars only. In total we have 14,108 companies (i.e., items) and 5,606 investors (i.e., users). In regards to investment relationships we have 40,544 relationships with an average of 7 relations per investor. Our most prolific investor has made slightly less than 400 past investments. On average a company has around 4 distinct relationships with investors.

Table 7.2: CrunchBase Dataset.

Attribute	Description
entity_id	Entity (i.e. company) identity
name	Entity (i.e., company) name
overview	Entity CrunchBase overview (i.e., textual description)
entity_permalink	Entity CrunchBase permanent link
financial_org_id	Financial organization (i.e., investor) identity
financial_org_permalink	Financial organization CrunchBase permanent link

VentureSource Dataset

VentureSource contains 21,610 investee private companies (i.e., items), 7,560 venture capital firms and 32,710 investment partners (i.e., two distinct sets of users). With regards to investments, VC firms have

83,264 and investment partners have 82,897 with an average of 11.01 and 2.53 past investments respectively. Our most prolific VC firm and investment partner have each, respectively, made 600+ and 60+ past investments. On average an investee private company has distinct relationships with 3.85 VC firms and 4.41 investment partners. Comparing the sparsity directly to other datasets, such as MovieLens1M [Gro13] (95.5%) and Netflix (99.8%), VentureSource is extremely sparse (over 99.9%) and long-tailed which will prove challenging for generating relevant recommendations using existing recommendation techniques.

Table 7.3: VentureSource Dataset.

Attribute	Description
e_entityid	Encrypted entity (i.e., company) identity
e_vcid	Encrypted venture capital firm identity
e_pid	Encrypted investment partner (i.e., individual) identity

We are interested in improving the accuracy and relevance of Top-N investment opportunity recommendation in our particular use case but also in evaluating the utility of alternative classification schemes (see Chapter 5). Through leveraging industry classification our content-based and hybrid (i.e., collaborative- and content-based) recommendation techniques have been developed. The two main approaches to recommendation are latent factor models and neighbourhood models [Kor08, KBV09]. We deploy these two distinct but well established approaches to this recommendation task: i) Neighbourhood methods, which analyse similarities between items or users, and ii) Latent factor models, which attempt to directly profile both users and items through matrix factorization. Our methodology seeks to demonstrate the utility of collaborative filtering (CF) and content-based models in the task of Top-N investment opportunity recommendation.

7.3.2 Neighbourhood Methods

The recommendation models we use are based on the item-based k-Nearest Neighbour (kNN) [DK04]. The reason for choosing item-based neighbourhood models is because (i) it is still the most frequently used recommendation method in industry applications; (ii) it naturally incorporates the company attributes such as industry hierarchy.

The k-nearest neighbour algorithm (kNN) is a method for classifying objects based upon the nearest training examples in the feature space. k-Nearest Neighbours [SM86] algorithm classifies input vectors based on a majority vote of its neighbours, assigning the vector to the most common class amongst its k nearest neighbours.

- kNN is trained using vectors of a d-dimensional feature space. This feature space is partitioned into regions and labels using a training set.
- A point (i.e., vector) in the space is assigned to a class if it is the most frequent class among the k-nearest training samples.

- During the test phase, the distances of a new vector to all stored vectors are computed and the k closest ones are selected.

- Then the most common class amongst the k neighbours is chosen and assigned to the new vector.

In our experiments cosine similarity has been used for calculation of the distances between the vectors. The key component in item-based models is the item-item similarity function. Specifically, we have two basic settings of the item-item similarity. The first one is based on the cosine similarity of the industry hierarchy of the companies (i.e., content-based). This is repeated using the various classification schemes as inputs for training the kNN recommender model. The second is based on the overlap of the VC firms or investment partners of investee companies (i.e., collaborative filtering). We develop various item-based models based on these two item-item similarity functions. Moreover, we also leverage the linear ensemble method [Zho12] to combine the advantages of the two different models in a hybrid model.

7.3.3 Latent Factor Models

As described by [KBV09] matrix factorization models map both users and items to a joint latent factor space of dimensionality f, such that user-item interactions are modelled as inner products in that space. Accordingly, each item i is associated with a vector $q_i \in \mathbb{R}^f$, and each user u is associated with a vector $p_u \in \mathbb{R}^f$. For a given item i, the elements of q_i measure the extent to which the item possesses those factors, positive or negative. For a given user u, the elements of p_u measure the extent of interest the user has in items that are high on the corresponding factors, again, positive or negative. The resulting dot product, $q_i^T p_u$, captures the interaction between user u and item i — the user's overall interest in the item's characteristics. This approximates user u's rating of item i, which is denoted by r_{ui}, leading to the estimate:

$$\hat{r}_{ui} = q_i^T p_u \tag{7.1}$$

The major challenge is computing the mapping of each item and user to factor vectors $q_i, p_u \in \mathbb{R}^f$. After the recommender system completes this mapping, it can easily estimate the rating a user will give to any item by using Equation 7.1. Such a model is closely related to singular value decomposition (SVD), a well-established technique for identifying latent semantic factors in information retrieval.

There are many methods for item recommendation from implicit feedback like matrix factorization (MF) or k nearest-neighbour (kNN). Even though these methods are designed for the item prediction task of personalised ranking, none of them is directly optimised for ranking. Since this is a Top-N item recommendation problem, we adopt the ranking-aware Bayesian Personalized Ranking (BPR) [RFGST09]

for our basic collaborative filtering model. As a typical pair-wise learning to rank model, BPR takes each pair of positive-negative items as the training unit and the learning goal is to correctly predict which one of the two is positive/negative. More details of BPR learning framework can be found in [RFGST09, ZCWY13]. Here we mainly focus on the item scoring modelling $r_{u,i}$ in BPR. For basic BPR model, the prediction $\hat{r}_{u,i}$ is

$$\hat{r}_{ui} = \mu + b_u + b_i + p_u^T q_i, \tag{7.2}$$

where p_u and q_i are the latent factor vectors of user u and item i, and μ, b_u, and b_i are global, user, and item bias terms respectively.

With the industry classification information for each company (i.e., item), we can refine the scoring model as

$$\hat{r}_{ui} = \mu + b_u + b_i + p_u^T \Big(q_i + \sum_{j=1}^{n} c_{i,j} \rho_j \Big), \tag{7.3}$$

where ρ_j is the latent factor for category j and $c_{i,j}$ is the value for the item i assigned to the category j as discussed in Eq. (5.1).

Furthermore, in order to improve the recommendation performance we can directly incorporate the description of companies into the model as

$$\hat{r}_{ui} = \mu + b_u + b_i + p_u^T \Big(q_i + \sum_{j=1}^{n} \phi(d_i)_j \varrho_j \Big), \tag{7.4}$$

where the function $\phi(d_i)$ maps the description d_i to a lower-dimensional space using PCA and $\phi(d_i)_j$ is its component on jth dimension, ϱ_j is the corresponding latent factor vector for this PC dimension. On the other hand, the item neighbourhood model (SVD++) [Kor] can be leveraged:

$$\hat{r}_{ui} = \mu + b_u + b_i + p_u^T q_i + |N(u,i;k)|^{-\frac{1}{2}} \sum_{j \in N(u,i;k)} w_{ij}(r_{uj} - \bar{r}_u), \tag{7.5}$$

SVD++ refers to a matrix factorization model which makes use of implicit feedback information. In general, implicit feedback can refer to any kinds of users' history information that can help indicate users' preference. In our scenario the past investment history of the VC firm or investment partner.

By combining different item information incorporated models together, we can obtain different informative recommendation models.

7.4 Experiments

7.4.1 Experimental Setting

VentureSource and CrunchBase are chosen as the test datasets in our experiment. Using recommendation techniques to recommend top-N investment opportunities, specifically an implementation of a content-based recommender system, we intend to compare different existing classification schemes (outlined in Section 4.1) and generated schemes (described in Section 4.2—4.5) as input features. We then compare and combine through a hybrid approach collaborative filtering techniques based purely on user-item interactions. Primarily using data provided by Dow Jones VentureSource, we have access to historical US financings since 1987. For each entity (i.e., private company) we have the relationships to venture capital (VC) firms and individual investment partners.

For the neighbourhood methods the MyMediaLite recommender system library [GR11] is used to implement an item-based k-Nearest Neighbour collaborative filtering (CF) approach. We incorporate both existing (Group, Segment, Code) and generated (Multi) item attributes (i.e., industry hierarchy) in order to improve our performance. We use RapidMiner (an open-source data mining environment) and e-LICO (an e-Laboratory for interdisciplinary collaborative research in data mining and data-intensive science). Using the e-LICO recommender extension [MAFv12] for RapidMiner it is possible to implement a content-based recommender system based upon the MyMediaLite recommender system library [GR11].

For the latent factor models previous work on [ZCWY13] is followed using item Top-N collaborative filtering using a random 4:1 train and test split on the data. For the compared CF algorithms, we have the Bayesian Personalized Ranking (BPR) [RFGST09] models incorporating the Full, Partial, and Multi item category information, and their description PC information. We combine the category models and the PC model, also we add the item neighbourhood setting (SVD++). Specifically, we implement these algorithms using the SVDFeature [CZL12] toolkit.

7.4.2 Evaluation Measures

It is possible to utilise the previously generated principal components disregarding classification schemes entirely. Providing a useful benchmark against the existing and other generated classification schemes. In order to assess the performance of our recommender system we have decide to benchmark our classification schemes against:

1. a random recommender and

2. using raw PC representation (i.e., no classification scheme)

We observe standard performance metrics based upon 4:1 split validation. In order to evaluate the performance of our recommendation model we calculate the following commonly used evaluation

metrics: area under the curve (AUC), precision at cut-off rank k (Prec@k) and mean average precision (MAP) [GR11].

Area under the curve (AUC)

Area under the curve (AUC) refers to area under the receiver operating characteristic (ROC) curve. ROC curves offer a two-dimensional graphical representation of classifier performance (i.e., a recommender system, an information retrieval system or any other type of binary classifier) [Faw06]. ROC curves plots recall (i.e., true positive rate) against fallout (i.e., false positive rate) for increasing recommendation set size.

As outlined by [STL11] a perfect recommender would yield a ROC curve that goes straight up towards 1.0 recall and 0.0 fallout until all relevant items are retrieved. Afterwards it would go straight right towards 1.0 fallout while the remaining irrelevant items follow. The obvious aim is consequently to maximise the area under the ROC curve. The area under the curve (AUC) can therefore be used as a single measure for the overall quality of a recommender system. However, because random guessing produces the diagonal line between (0,0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5 [Faw06]. Figure 7.4 illustrates the AUC performance of our kNN recommender systems experiments at varying values of k and using different techniques such as collaborative filtering (e.g., itemKNN CF) and content-based, using different levels of the industry classification hierarchy (e.g., Group, Segment, Code).

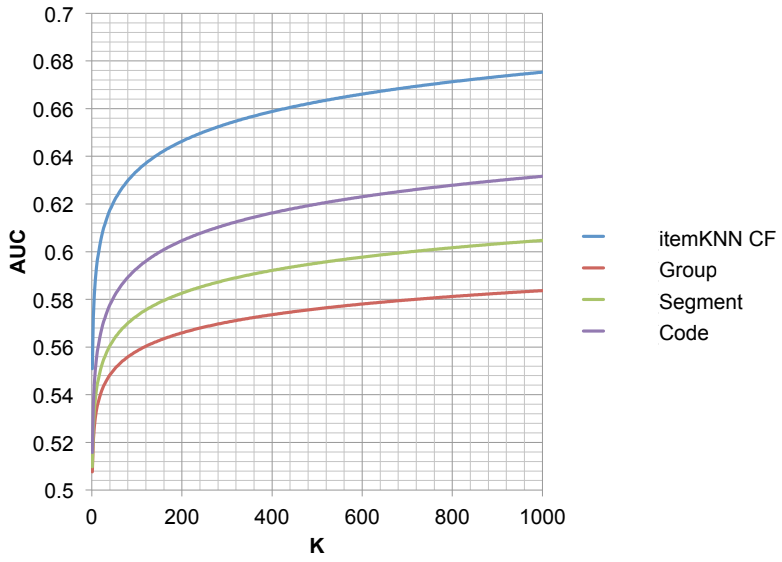


Figure 7.4: AUC Performance of Different Input Industry Classification Schemes Using kNN at Varying k on VentureSource dataset.

e_vcid	e_entityid	rank
690099	155053	1
690099	138053	2
690099	165053	3
690099	150053	4
690099	148053	5
690099	176053	6
690099	139053	7
690099	180053	8
690099	140053	9
690099	136053	10
690099	181053	11
690099	141053	12
690099	146053	13
690099	167053	14
690099	144053	15

Figure 7.5: Ranked top-k (k=15) Recommendations for VCID=690099.

Precision@K (prec@k)

Precision calculated at a given cut-off rank k, considering only the top k most results returned by the system and ignoring those documents ranked lower than k. Whilst precision at k is a useful metric (e.g., prec@10 corresponds to the number of relevant results on the first search results page), but fails to take into account the ranking of the relevant documents among the top k.

In the figure 7.5 we have retrieved top 15 recommendations for a particular investor. Observing our dataset we find that this investor (e_vcid=6900099) has previously invested in two different companies (e_entityid=[131053, 139053]). Precision is the number of relevant items retrieved divided by the number of retrieved items, therefore, our precision metrics would be calculated as follows: prec@5(0/5=0) 0%; prec@10 (1/10=0.10) 10%; prec@15(1/15=0.067) 6.7%; MAP(0.868/9=0.0965) 9.65%. These precision metrics (similarly for other metrics) are then averaged over the results for every user in order to obtain our recommender system performance metrics.

Mean average precision (MAP)

As noted by [HKTR04] and others ([MRS08], [STL11]) mean average precision (MAP) is a popular metric for search engines and is applied, for example, to report results at the Text REtrieval Conference (TREC) [Nat09]. It takes each relevant item and calculates the precision of the recommendation set with the size that corresponds to the rank of the relevant item. Then the arithmetic mean of all these precisions is formed. Afterwards we calculate the arithmetic mean of the average precisions of all users to get the final mean average precision.

As defined by [MRS08] average precision (AP) is the average of the precision value obtained for the

set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs. That is, if the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \ldots d_m\}$ and R_{jk} is the set of ranked retrieval results from the top result until we get to document d_k, then:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_j k)$$

For MAP@K, the goal is to evaluate the average precision of the predicted recommendation list for each user. Suppose a user has n item interactions in the test dataset and the system can recommend up to K items to this user. The average precision score at K (ap@K) is: $ap@K = \sum_{k=1}^{K} P(k)/min(n, K)$ where $P(k) = 0$ if the user has no interaction the k-th product of the recommended list in the test dataset and $P(k) = k$ if otherwise. The mean average precision for M user at K (MAP@K), is the average of the average precision of each user, which is defined as: $MAP@K = \sum_{m=1}^{M} ap@K/M$. The higher the MAP@K score, the better the recommender system performs.

7.4.3 Experimental Results

Neighbourhood Methods

The overall results of the compared algorithms are shown in Table 7.4 for VC firms and Table 7.5 for investment partners. From our results we make the following observations: (i) All the algorithms obtain improved performance against the baseline Random, which indicates the efficacy of our item-based CF and ensemble models. (ii) The recommendation performance is improved by introducing the existing industry hierarchy information (Group, Segment, Code). (iv) By combining the item-based kNN CF and industry Code information using the linear ensemble method, we get our empirical best model on the AUC measure. (v) Initially the multiple industry assignment (Multi) leads to some additional improvement, however, it seemingly has no significant impact on the ensemble methods. (vi) The values of MAP and precision are quite low, which is most likely due to the extreme sparsity of the VentureSource dataset.

Table 7.4: Performance of kNN on VentureSource Dataset for VC Firms.

Model	AUC	Prec@5	Prec@10	Prec@15	MAP
Random	0.4999	0.0000	0.0001	0.0002	0.0006
Group	0.5590	0.0008	0.0007	0.0006	0.0016
Segment	0.5700	0.0012	0.0011	0.0008	0.0016
Code	0.5920	0.0013	0.0013	0.0014	0.0023
Multi Group	0.5727	0.0006	0.0006	0.0006	0.0014
Multi Segment	0.5730	0.0008	0.0008	0.0008	0.0016
Multi Code	0.5841	0.0012	0.0011	0.0096	0.0023
CF	0.6362	0.0127	0.0099	0.0083	0.0169
CF + Group	0.6430	0.0129	0.0101	0.0087	0.0172
CF + Segment	0.6477	0.0131	0.0107	0.0090	0.0185
CF + Code	0.6582	0.0143	0.0108	0.0091	0.0175
CF + Multi Group	0.6440	0.0125	0.0100	0.0083	0.0169
CF + Multi Segment	0.6414	0.0113	0.0094	0.0079	0.0153
CF + Multi Code	0.6478	0.0126	0.0096	0.0081	0.0165

Table 7.5: Performance of kNN on VentureSource Dataset for Investment Partners.

Model	AUC	Prec@5	Prec@10	Prec@15	MAP
Random	0.4947	0.0001	0.0001	0.0001	0.0007
Group	0.5557	0.0001	0.0001	0.0001	0.0009
Segment	0.5687	0.0002	0.0002	0.0002	0.0013
Code	0.5825	0.0004	0.0005	0.0005	0.0018
Multi Group	0.5604	0.0001	0.0001	0.0002	0.0009
Multi Segment	0.5663	0.0004	0.0003	0.0003	0.0015
Multi Code	0.5783	0.0005	0.0004	0.0005	0.0019
CF	0.6163	0.0093	0.0069	0.0057	0.0203
CF + Group	0.6233	0.0094	0.0071	0.0058	0.0210
CF + Segment	0.6283	0.0092	0.0069	0.0056	0.0197
CF + Code	0.6312	0.0087	0.0063	0.0051	0.0188
CF + Multi Group	0.6216	0.0089	0.0067	0.0055	0.0199
CF + Multi Segment	0.6227	0.0085	0.0062	0.0051	0.0183
CF + Multi Code	0.6234	0.0071	0.0054	0.0045	0.0161

Latent Factor Models

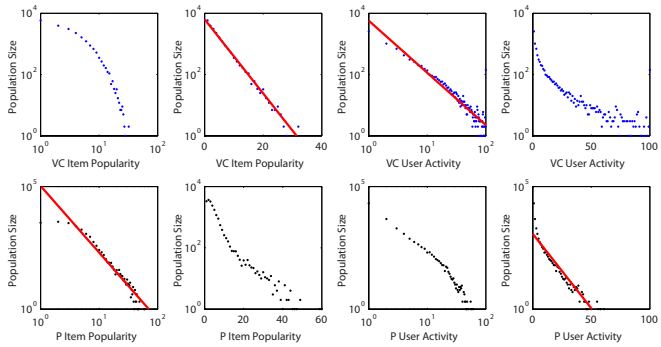
Latent factor models appear not as effective as the neighbourhood methods, which might be caused by the extreme sparsity and unique user-item interaction properties. The results of the compared algorithms on the CrunchBase dataset are shown in Table 7.6. From our results we make the following observations: (i) All the algorithms obtain improved performance against the baseline Item-kNN, which indicates the efficacy of introducing classification and CF models. (ii) The general performance on AUC is not as satisfactory as traditional CF datasets (such as 0.92 on Netflix [RFGST09]), which indicates the difficulty of performing traditional CF algorithms on this investment opportunity recommendation task. (iii) The recommendation performance is improved by introducing the industry classification information (Full, Partial, Multiple), company description information (PC), and company neighbourhood information (SVD++) respectively. (iv) By combining the Partial and PC information in the setting of SVD++, we get our empirical best model on the AUC measure. (v) The values of MAP, precision, and recall are quite low, which is due to the extreme sparsity of the CrunchBase dataset.

Table 7.6: Performance of Informative CF Models on CrunchBase Dataset.

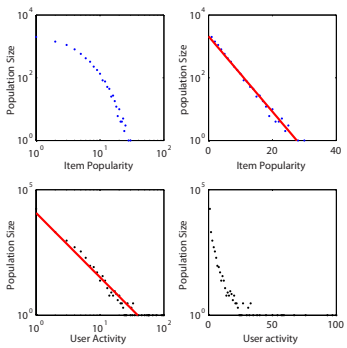
Model	AUC	MAP	Prec@5	Rec@5
BPR	0.7343	0.0037	0.0014	0.0034
BPR Full	0.7827	0.0062	0.0027	0.0054
BPR Partial	0.7321	0.0036	0.0011	0.0026
BPR Multi	0.7888	0.0051	0.0016	0.0040
BPR PC	0.7483	0.0040	0.0014	0.0030
BPR Full PC	0.7943	0.0065	0.0030	0.0055
BPR Partial PC	0.8091	0.0046	0.0014	0.0024
BPR Multi PC	0.8009	0.0050	0.0016	0.0029
SVD++	0.7461	0.0041	0.0020	0.0035
SVD++ Full	0.7829	0.0070	0.0027	0.0058
SVD++ Partial	0.7413	0.0040	0.0019	0.0035
SVD++ Multi	0.7914	0.0055	0.0023	0.0035
SVD++ PC	0.8094	0.0052	0.0020	0.0034
SVD++ Full PC	0.7946	0.0060	0.0030	0.0057
SVD++ Partial PC	0.8178	0.0050	0.0022	0.0040
SVD++ Multi PC	0.8028	0.0052	0.0025	0.0038

7.5 Discussion

We have explained our motivation for applying recommender system to venture finance and our specific intended use cases. Furthermore, we have developed methodology and a variety of techniques for both i) identifying peers through company similarity measures (in the previous Chapter) and ii) Top-N



(a) User-item interaction difference between venture capital firms (VC) and investment partners (P).



(b) Simulation of user-item interaction.

Figure 7.6: User-item Interaction Analysis.

investment opportunity recommendation. There is a question of the "material impact" of such recommendations with our current implementation. We have input investment history plus industry vectors of entities and output recommendations for each VC (or partner). However, the performance metrics are relatively poor (e.g., MAP = 0.01; AUC = 0.55) and although there is an improvement through using different industry vectors (i.e., lower level of hierarchy (industry segment, industry code), multiple assignment) as inputs it is only incremental (e.g., MAP = 0.02; AUC = 0.57). On the other hand using collaborative filtering (CF) techniques we see slightly more reasonable metrics (e.g., MAP = 0.02; AUC = 0.63). The main issue with CF techniques is the "cold start" problem where we have no prior information with which to infer recommendations for new users (i.e., VC or partner). Ideally, we need external validation of our recommendation models which is why undertaking a further user study with industry is necessary.

In order to better understand the performance and challenges of generating recommendation on both the VentureSource and CrunchBase datasets we now will discuss investors' specialisation by industry (as discussed previously in Section 7.2), the motivation for utilising industry classification in content-based techniques and user-item interactions, especially compared to other domains, which should provide insight into the effectiveness of collaborative filtering (CF) models.

7.5.1 User-Item Interaction Analysis

We conducted an investigation of the user-item interaction data from VentureSource, observing both the VC firm and investment partner interactions with investee private companies. In Figure 7.6, we show histograms with log-log and semi-log of the user and item popularity distributions identifying some interesting characteristics.

For the VC firm and investment partner datasets, the distributions appear to be quite distinct. Observing Figure 7.6, in relation to item popularity, the VC firms follow an exponential distribution (upper second panel), while the investment partners follow a power-law (lower first panel). On the contrary, for user popularity, the VC firms follow a power-law (upper third panel) and investment partners follow an

exponential distribution (lower fourth panel).

From this observation we can see that the two datasets have some fundamental differences in their network properties, which may coincide with the algorithm performances on the two datasets. From a complex network perspective, the power-law distribution represents the existence of very active VCs, which is referred to as "the fat tail effect" [AH00]. In comparison, the exponential distributions on the investee private companies and the investment partner sides indicate that there is no such extreme properties. Since each investee private company or individual partner cannot have such a large number of investment relationships (i.e., there is a natural limit to how many investors invest in a single company).

In order to explain such network properties, specifically from the VC firm perspective, we ran a simulation (see Figure 7.6(b)). For items (upper two panels), it displays an exponential distribution. The rule is that the probability that one item gathers one more connection is proportional to its current degree. For users (lower two panels), it displays a power-law distribution. The rule is that the probability that one user creates one more connection is proportional to the second order of its current degree. Hence, the effect that the "rich get richer" is even greater in approximating the power-law distribution (i.e., for the VC firm). This corroborates with much of the relevant financial literature on VC networks and superior performance [SHH99, LLP01, SS01, YRAS01, SC02, Sor03].

In the simulation experiment, if we sample the user/item by their current degrees, both sides will have exponential distributions. In order to model the power-law distribution, we need to be more biased on the node degrees. As a result, when we sample one side by a quadratic form of the nodes' current degree, we can approximate the power-law. An explanation might be that active VC firms (i.e., those that making more investments) become more popular and well known subsequently receiving more investment opportunities and therefore making more investments. From the perspective of the VC firm, if VC firm A has two times as many investments as VC firm B, then A is more than two times popular than B, which makes the investment popularity of VC firms a power-law distribution.

Comparison To Other Domains

We conduct an investigation of the user-item interaction data from CrunchBase by comparing it with Netflix, a well-known CF movie benchmark dataset. First, in terms of data sparsity, the observed rating ratio of Netflix is only 1.17×10^{-2}, while that of CrunchBase is even lower: 5.13×10^{-4}, which is 22 times lower than Netflix. Such sparsity is reasonable since private investment activity is not as commonplace as simply watching a movie. The final investment decision will require the consent of both the company and investors usually involving a lengthy due diligence process. Furthermore, in Figure 7.7, we obtain the histograms with log-log and semi-log of the user and item popularity distributions and identify some interesting characteristics.

For the Netflix and CrunchBase datasets, the distributions appear to be quite distinct. Observing

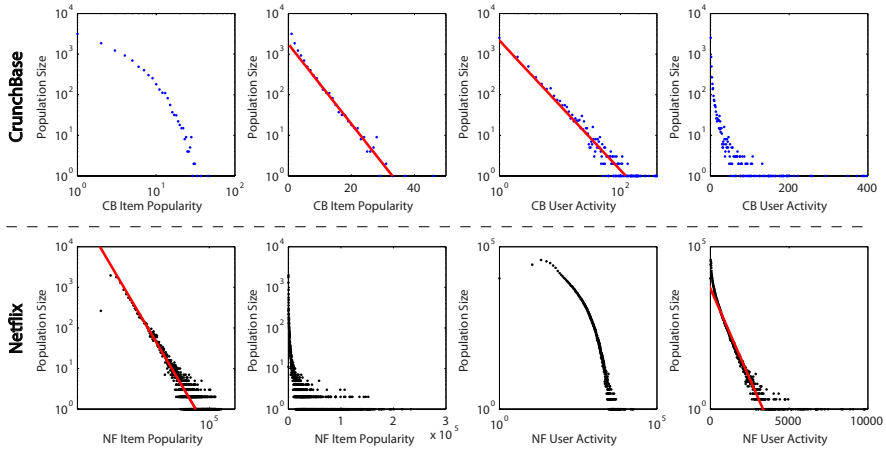


Figure 7.7: User-item Interaction for CrunchBase and Netflix.

Figure 7.7, for example, for the item popularity distribution, CrunchBase follows an exponential distribution (upper second panel), while the Netflix item popularity follows power-law (lower first panel). On the contrary, for the user popularity distribution, CrunchBase follows power-law and Netflix follows exponential. From this observation we can see that the two datasets have some fundamental differences in their network properties, which may coincide with the algorithm performances on the two datasets. From a complex network perspective, if a distribution looks like power-law, it maybe due to the fact the connections are established in a way that the "rich get richer" [AH00]. For Netflix, people tend to watch frequently rated or discussed movies, which makes the popular movies even more popular. While for CrunchBase, the user (i.e., investor) activity follows power-law instead of the items (i.e., companies), which may indicate that the well-known investors are more likely be offered an investment opportunity (commonly referred to by investors as "deal flow") in the private companies seeking investment. We could infer that there may be a compounding effect whereby companies seek investment from the most popular (or well known) investors (e.g., Sequoia Capital, Intel Capital). However, other factors such as fund size, location and investment strategy would likely also have an effect.

However, we are artificially creating a train and test split in order to generate such "offline" performance metrics for our recommendation models. In reality, many investment opportunities will be relevant or of interest to a particular investor, not simply those in which they have made past investments. Therefore there is an onus to undertaken further evaluation of our recommendation techniques in a real-world industry scenario.

We have demonstrated the application of recommendation techniques in relation to the novel application domain of venture finance. Through our venture finance data analysis, we discover fundamental differences in user-item interaction patterns between VC firms and their individual investment partners. Our methodology takes advantage of our access to venture financing data to improve the investment opportunity recommendation quality on the VentureSource and CrunchBase datasets. Whilst we em-

pirically shown (i.e., AUC, MAP) an improvement in utilising both existing and generated industry classification schemes in our recommendation strategies further user studies with industry are required to validate the real-world benefits of such methods.

Part IV

Conclusions

Chapter 8

Conclusions

This concluding chapter provides a discussion and critical assessment of the research undertaken including contributions, limitations, experience working with industry and further work.

8.1 Contributions

This research has focused on improving upon existing industry classification schemes and applying recommender systems to venture finance. We have generated a novel form of industry classification using multi-label classification and, in collaboration with Correlation Ventures, demonstrated real-world application of recommender systems to the domain of venture finance. Our improved industry classification scheme addresses several of the shortcomings of existing industry classification schemes and has been utilised for relevant use cases in venture finance. Including peer identification through refining similarity measures, and improving the performance of recommendation models in the task of Top-N investment opportunity recommendation. Finally, a pilot user study was undertaken in to objectively test and compared different classification schemes.

Additionally, the design, testing and implementation a web-based tool, NVANA (New Venture Analytics) in collaboration with various partners for assessing new ventures. Our prototype and the potential to extend such a system based upon work on classification and applying recommender system techniques to this domain serve as a novel contribution.

8.2 Scope for Computational Analytics

Table 8.1 below provides a detailed but non-exhaustive list of existing computational analytics with applications in venture finance based upon our survey of software tools (see Chapter 3):

Table 8.1: Current Applications of Analytics for Early-stage Investment.

Stage	Computational Method	Application	Example
Deal Origination Screening	Classification	Venture Classification (Stage, Required investment size, Industry, Location) Investor Classification (Stage, Investment size, Industry, Location)	VentureSource CrunchBase AngelList PitchBook CB Insights
Evaluation	Regression	Scoring & Ranking Due Diligence	Growth Intelligence CB Insights Mosaic Mattermark Inkwire Bright*Sun Datafox Indicate.io
Structuring	Simulation	"What if" analysis	Solium CapMx
Post investment activities	Classification	Benchmarking	Dashboard.io Compass

8.2.1 Stage of VC Investment Process

Deal Origination & Screening

In relation to Deal Origination, the processes by which deals enter into consideration as investment prospects, there is seemingly scope for applying analytics to Investment Networks, Databases and Market Intelligence platforms. In the section on Analytics we have observed how classification and recommender systems techniques (i.e., similarity, top-N recommendation) can be applied. There is little reason why these could not be successfully deployed in these types of tools.

In relation to Databases the use of auto-classification via supervised learning would remedy the issue of unclassified or misclassified companies and some of the short falls of overly simplistic categorisation used by some databases. Given the appropriate training set and textual data it is feasible to reclassify companies against any preferred scheme (e.g., VentureSource) and also to generate multi-class industry representations (e.g., $I_{Multiple}$, $I_{Partial}$). As discussed in Chapter 5 there is also scope for multi-dimensional classification schemes (e.g., Stage, Business model, Customer type, etc) beyond purely industry classification.

For both Investment Networks and Market Intelligence platforms classification techniques described have the potential to improve screening based upon relevant industries and for identifying similar companies as demonstrated using datasets from AngelList [1]. Furthermore, we open the possibility of automating the matching of companies and investor through applying recommender systems. Either using basic collaborative filtering methods or by using pertinent content-based methods (e.g., industry, stage, business model, etc). Observing the number of companies currently listed on these platforms CB Insights tracks 50,000 VC backed companies [Ins13], AngelList currently totals 79,389 [Ang13b], CrunchBase stats count 197,470 companies [Cru13b] and Mattermark claims 200,000 startups [Mat13]. Whilst not all active this is not an insignificant number of companies for a VC to keep track of. Therefore, it is understandable why they my deploy existing screening tactics such as relying on referrals from trusted

[1] `http://angellist-demo.prediction.io/`

sources (entrepreneurs, lawyers, accountants, other investors).

Beyond the use of classification and recommender systems techniques there is also scope for estimating private company performance to be determined using regression analysis as described in Chapter 6 further supporting the screening of companies. As discussed in Chapter 3 there are more and more tools focusing in this area by providing indices or scores for private companies (e.g., CB Insights Mosaic, Mattermark, Inkwire, Datafox, SignalFire, Bright*Sun etc).

Whilst there is seemingly potential for improving the screening capabilities of a VC investor the more in depth due diligence and evaluation is most likely still as very manual or human undertaking perhaps informed by data from new tools.

Evaluation & Structuring

In relation to Evaluation, the assessment of perceived risk and expected return on the basis of a weighting of several characteristics of the prospective venture and the decision whether or not to invest as determined by the relative levels of perceived risk and expected return. There is seemingly less scope to analytics in terms of Deal Platforms for investors to manage their deal flow of prospective investments and Specialised CRM Systems used in later-stage private equity.

Of the numerous tools covered by our survey (see Chapter 3), Younoodle, which no longer exists in its previous form, is certainly of interest. Younoodle attempted to offer a forward-looking measure of future enterprise value based on historical returns and tracking real-time data associated with a company. However, there is obviously difficultly in capturing all relevant variables to accurately predict the future value of a company. After all, past performance is not a reliable predictor of future performance. Although investors and industry commentators were heavily skeptical and Younoodle now ceases to exist (now merely a entrepreneurship-focused social network and competition platform alongside spin-off company Quid, focusing more on network analysis) we may be seeing history repeating itself. Even today we are seeing a more data-driven, or at least data-informed, approach to deal evaluation being pursued by members of the venture capital investment community. Predicting a company's future financial value is clearly difficult, maybe impossible. Largely due to the lack of operating history for the companies being evaluated. As such, evaluation is often a subjective judgment based on a multidimensional set of characteristics [TB84]. Whilst some venture capitalists, especially those focusing on the later stages of investment, attempt to calculate risk and return this is not always feasible with start-ups and *early-stage* investment due to the inherent uncertainties involved.

Structuring, the negotiation of the price of the deal, namely the equity relinquished to the investor, plus the covenants which limit the risk of the investor. The application of analytics to Structuring and Capitalisation Management tools is fairly limited beyond scenario-based "What if" analysis. Arguably, a multi-label classification scheme may be useful in identifying similar companies for comparables in

the structuring and valuation of private companies. However, there are usually few relevant private (e.g., other recent financing events) or even public company comparables (e.g., recent IPOs or publicly traded companies) for any particular company at any particular stage of investment. In the case of private financing the valuation may not necessarily even be public information.

Post-investment Activities

Although not the focus of the rest of this research, the broadness of Post-investment Activities covering software assistance to the venture in the areas of recruiting key executives, strategic planning, locating expansion financing, and orchestrating a merger, acquisition or public offering. Therefore, within the scope of analytical methods, for example with Benchmarking tools and secondary Markets for private company shareholdings. Again techniques discussed shortly in relation to supervised classification and recommender systems are both pertinent. For example, in benchmarking private companies it is important to identifying relevant peers. Estimating performance, such as revenue, via some proxy measures is less likely to be valuable at this stage as it is assumed the investor has readily available access to *actual* performance data. Perhaps, there is still value in estimating performance of other private companies (i.e., competitor analysis) in relation to a particular investment.

8.2.2 Stage of Investment

Beyond the stage of the investment process (i.e., origination, screening, ...) we can also look at the scope of applying analytics based upon the stage of investment (i.e., seed, start-up, expansion). Clearly the stage of investment is an import factor in understanding the potential value in applying analytics. It would seem in terms of "pre-seed" stage, also referred to as "concept" or "idea" stage when a company is pre-product then there is seemingly very little data with which to perform any meaningful analysis, especially in terms of traction but also in relation to product or market. In fact the only relevant data might be past track record of information about the founding team. It is no surprise then that the team is commonly deemed the most important factor in making early-stage investment decisions. However, even this is contentious with others suggesting a bad team can be remedied or replaced [KSS09] and importantly different investors have different profiles (e.g., people, technology, financial) [CKL05b].

In Figure 8.1 we illustrate the scope for analytics superimposed over the traditional startup financing cycle [Sta12] and the different sources and stages of funding in relation to time and revenue. Different analogies for funding include, "gears" [Gra05] or "ladders" [oT07], to convey the fact that ventures often go through several sequential rounds of funding. Investment of risk capital can also be segmented into stages of funding. The British Venture Capital Association (BVCA) commonly defines the stages as early-stage, growth and management buy-out (MBO) or buy-in (MBI) [BVC11]. Although definitions vary, "early-stage" is generally perceived to be less than £2m. Different types of investors will look to invest at different stages of investment; those associated with early-stage funding generally include

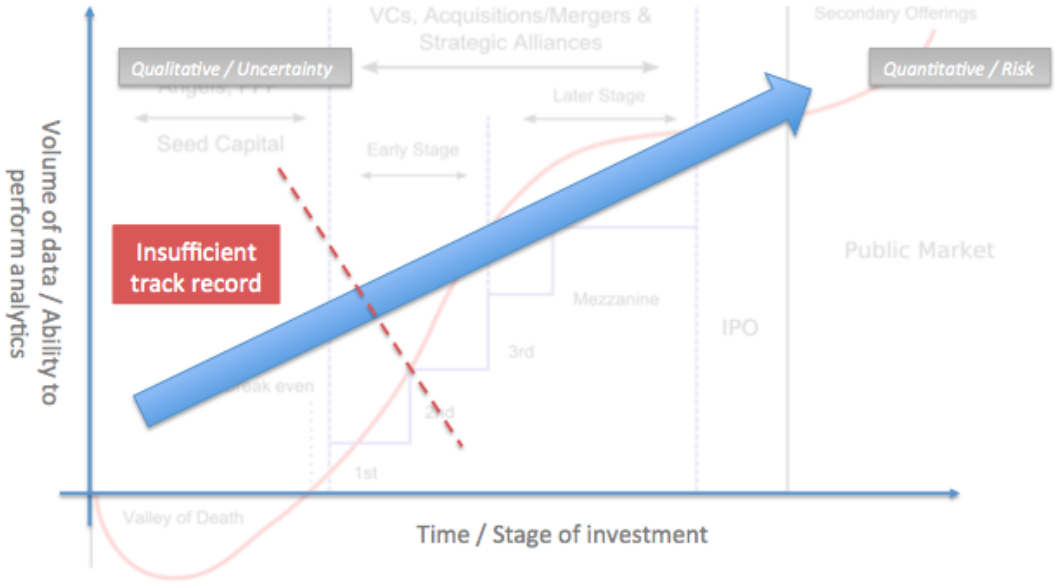


Figure 8.1: Scope for Analytics by Stage of Investment [Sta12].

public or government agencies, universities, seed accelerators, angel investors and *some* venture capital firms. Clearly, other sources of finance such as organic growth (e.g., sales revenue) and debt financing (e.g., bank loans) have not been mentioned, this is because there are not typically assumed to be forms of "risk capital", however, they often provide necessary financing for many businesses.

Defining the appropriate stage of investment for applying analytics is currently somewhat arbitrary, it is likely to be more relevant to "Series A" stage of investment (i.e., the first round of institutional capital) and beyond, once the company has some history of revenues and track record. The ability to perform meaningful analysis is most likely to be dictated by the data availability (e.g., Filling accounts with Companies House; AlexaRank less than 100,000; etc.). Seemingly the "pre seed" stage at which NVANA was implemented and tested is not particularly well suited to collecting useful data or applying meaningful analytics due to the fact that the new ventures being assessed are at the concept or idea stage and often pre-product. Whilst focusing on later stages of investment (i.e., "Expansion" or "Late" stage) would lead to an increased availability of relevant data both financial and non-financial indicators there is arguably less of a necessity for analytics and automation due to the decreased volume of prospective companies at this stage.

Some may argue that it is still possible to collect useful data at this very early "pre-seed" stage but the important indicators are not the same as later stage companies (i.e., financials) but instead should be focused on initial validation of an idea or product. For example, focuses on number of customer conversations and validated hypotheses into an "investment readiness" score [Bla13] and also other relevant factors not related to the product or traction such as the team members' background and experience. Po-

tentially the "sweet spot" for effectively applying analytics to venture finance is somewhere in between in the "start-up" stage whereby the is a large number of prospective companies leading to the need for automation and a reasonable track record or time period in which to collect relevant data points on which to perform meaningful analysis regarding a company's performance and future potential.

Ultimately the goal and objective for a VC is to maximise returns on their fund (i.e., measured by IRR or multiples). Whilst we have seen the performance of VC funds is high variable, at least in theory, the way to achieve this is by selecting "winners". Everything else from the structure of the fund, to the specific investment terms and valuation, to post-investment activities is secondary. If a VC fund fails to invest in companies which successfully exit and create a return for its limited partners they will simply cease to exist in the long run (i.e., they will be unable to raise additional funds).

A simplified model and real-world example of such fund-level economics conveys the assumptions and expectations made:

Assumptions:						
Fund Size	$ 100,000,000					
Term	10	years				
Mgmt Fee:	2.50%	in first 4 years				
	2.25%	in year 5				
	2.00%	in year 6				
	1.75%	in year 7				
	1.5%	in years 8, 9 & 10				
Carry	20%	of gains net of mgmt fees				
Average Initial Investment	$ 2,088,235	$1mm concept, $2.5mm trial, $3.5mm revenue				
Average Follow On Investment	$ 2,500,000	for concept stage and trial stage investments				
	$ 3,500,000	for revenue stage investments				
Average Total Investment	$ 5,300,000	per deal				
Total Deals	15					
Initial Investments Per Year	3	Year 1				
	4	years 2 & 3				
	3	year 4				
	1	year 5				
Winners	5	33%				
Money Back	5	33%				
Losers	5	33%				
Rounds Per Investment:	1	for loser				
	2	for money back				
	3	for winner				
	4	for concept stage winner				
Deals By Stage:	4	concept	2 losers, 1 money back, 1 winner			
	7	trial	2 losers, 2 money back, 3 winners			
	4	revenue	1 losers, 2 money back, 1 winner			
Average Return Multiple	-	for loser				
	1.25	for money back				
	6.5	for winner				
Average Holding Period (Yrs)	6	concept				
	5	trial				
	4	revenue				
	2	loser				

Figure 8.2: VC Fund Economics Assumptions.

The table shown in Figure 8.2 shows a model of how fund economics and returns are expected to be realised for a $100M fund managed by Union Square Ventures, a prominent East coast VC firm. Apart from covering some of the detail of a VC firm's business model (e.g., Management fees, Carry) it also provides useful insights into some of the key assumptions and expectations around number of deals, investment size, returns distribution and holding period for a $100M fund.

We have already covered the typical VC investment process. In terms of impacting returns there are

seemingly only a few possibilities which will drive returns:

1. Increase quality of prospective deal flow (i.e., "top of funnel")
2. Increase quality of investments (i.e., "bottom of funnel")
3. Increase odds/chances of successful outcomes (i.e., incremental increasing returns via structuring (i.e., pricing/valuation) or post-investment activities (i.e., recruitment, strategy, future financings))

These possible options relate to the stages of i) Deal Origination, ii) Screening & Evaluation and iii) Structuring or Post-investment Activities respectively.

However, if we now observe historical distributions of returns in Figure 8.3 from a VC portfolio we see how returns are really driven by outliers those small number of investments which return. With around $>25\%$ and $>60\%$ of value by investments which return 2-5 times and 5+ times respectively. This would suggest that the real focus for a venture capitalist must be identifying and investing in the small number of potential opportunities which can drive returns on their fund. It is important to note that the expected venture capital returns are not normally distributed but instead resemble more of a log-normal or log-levy distribution [Ogu13].

Based on this knowledge it makes sense for a VC investor to focus their efforts on origination and selection (i.e., points i) and ii) above) as opposed to optimising the subsequent structuring and post-investment activities (i.e., point iii) above). Taking the view that the impact of a VC investor as board member or otherwise is not a primary factor in the outcomes of a portfolio company. After all, history shows that the returns are driven by outliers. That is not to say that these activities (structuring, post-investment activities) are entirely irrelevant or not important but in terms of impacting the returns to a particular fund they are subordinate to deal origination and investment selection.

In recent years, as discussed in Chapter 2, we are seeing a trifurcation whereby [Ant12] we see an industry with i) top-tier firms (such as Sequoia Capital), ii) incubators and accelerators (such a Y Combinator), iii) and finally, those firms taking a more quantitative approach to funding (such as Correlation Ventures). Consequently this has led to a "barbell effect" whereby there are top-tiers funds seemingly raising very large size funds (i.e., \$500M+) and an influx of smaller incubator, accelerator and "super angel" or "micro VC" funds (i.e., $<$\$100M) with fewer funds occupying the middle ground. Clearly different VC firms have different strategies (i.e., by industries, fund size, geography, stage) and return profiles. In particular fund size has an impact of the venture fund "economics" and investment decisions.

More and more VCs are hoping to make increased use of data and applying computational analytics to inform investment decision making. It is important to note that, with the average holding period for VC investment is 6-8 years, it is not possible to predict the future outcomes of investment ex ante. That doesn't mean using data and analytics to better inform human judgement is futile. In fact many critics

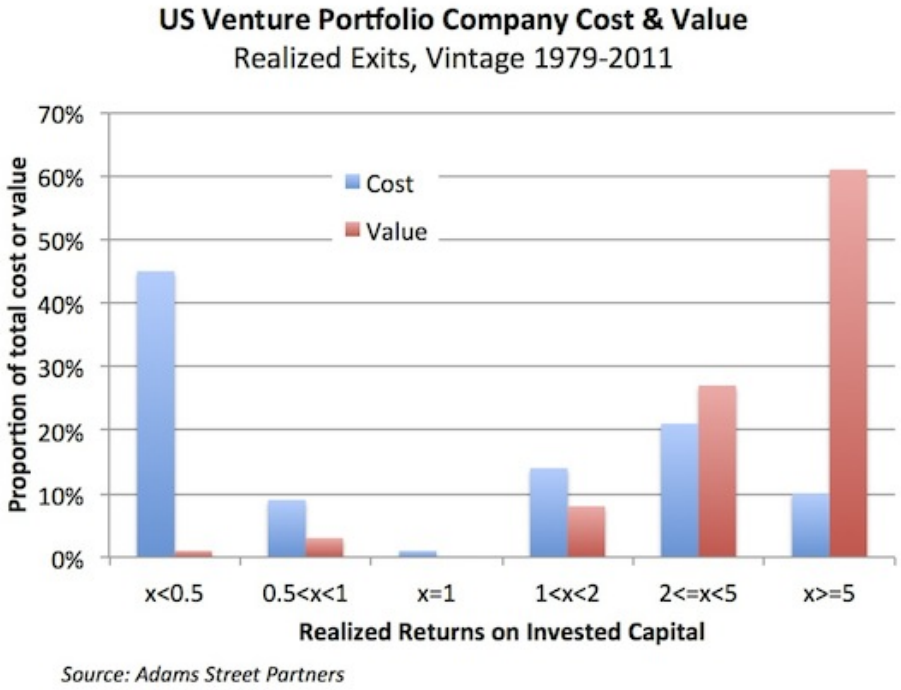


Figure 8.3: US Venture Portfolio Returns on Invested Capital.

of quantitative approaches to early-stage investment who cite that it is simply not possible to model the important factors (team, market, etc) are creating a false dichotomy of using data and algorithms versus human intuition, expertise and "gut instinct".

In fact, there are numerous potential applications of analytics to the venture finance process, which albeit not predict future outcomes of success or failure can definitely impact the performance of a VC firm especially when we consider increased competition and globalisation of the venture capital industry.

8.3 Real-world Application & Industry Collaborations

In order to validate a classification scheme we needed to define a mechanism by which we can compare and contrast the efficacy of different schemes. How do we judge what constitutes an improved classification scheme? In attempting to measure the efficacy of a classification scheme we need some perspective and context on the intended use case of such a scheme. It makes sense at this point to reiterate why we are interested in classification of private companies. Essentially we are taking the perspective of an investor. Whilst there are several alternative applications of an improved classification scheme (e.g., Matching companies with potential investors; competitor analysis; benchmarking peer-group performance; selecting valuation comparables; customer identification; market segmentation) our focus has been the following use cases:

- **Estimating Private Company Performance** — in order to analyse competitors and compare

potential investment opportunities by benchmarking company performance against a relevant peer-group of private companies

- **Top-N Investment Opportunity Recommendation** — improve the performance of recommendation models in the task of matching investors and companies. Also, for the purpose of finding co-investment partners and identifying the most suitable investors for a company and vice versa based upon past investment history.

Such applications offer potential value both to investors and private companies. In the case of peer-identification it allows both investors and private companies to identify potential competitors. In the case of matching it allows investors to find co-investors and relevant investment opportunities whilst allowing private companies to find relevant investors. Our intention was to utilise recommender systems as a basis for evaluating the utility of different classification schemes against these applications in venture finance (see Chapter 5). To some extent an improved classification scheme addresses the issues highlighted previously with existing schemes. However, we wish to empirically measure, test and compare different classification schemes beyond purely offline evaluation metrics or subjective judgement, which has been achieved through our user studies and collaboration with industry.

8.4 Further Work

We hoped to apply quantitative methods to early-stage investment akin to the manner in which banks had utilised credit scoring models to approve or reject loans. Unfortunately, whilst there is some overlap, the two problems are in fact quite different. Given the average holding period for a typical VC investment is 7-8 years before exit, predicting outcomes is no more possible than predicting future stock prices in the public markets. However, data about prospective investments current performance, the market position and other players is still extremely valuable. In fact there is much potential for more data-informed early-stage investment.

Whilst we have shown it is possible to improve upon existing industry classification schemes there is still potential for further improvement. In relation to classification schemes, multi-dimensional schemes (i.e., not purely industry focused) present an interesting area for future research. In relationship to the application of recommender systems to venture finance some suggestions include looking at potential return on investment as an overall objective as opposed to purely relevance in relation to Top-N investment opportunity recommendation.

Ultimately the perception that early-stage investment decisions should be undertaken using data, analytics and algorithms *or* through human judgement in isolation is a false dichotomy. As in many other domains, the application of machine learning is best complemented by human intuition. In fact, with regards to venture finance and early-stage investment decision making it would seem that data

informed decisions would be the best of both worlds. Only time will tell whether those VC firms applying quantitative methods and approaches to early-stage investment will perform better in the long run.

Part V

Appendices

Appendix A

Glossary of Terms

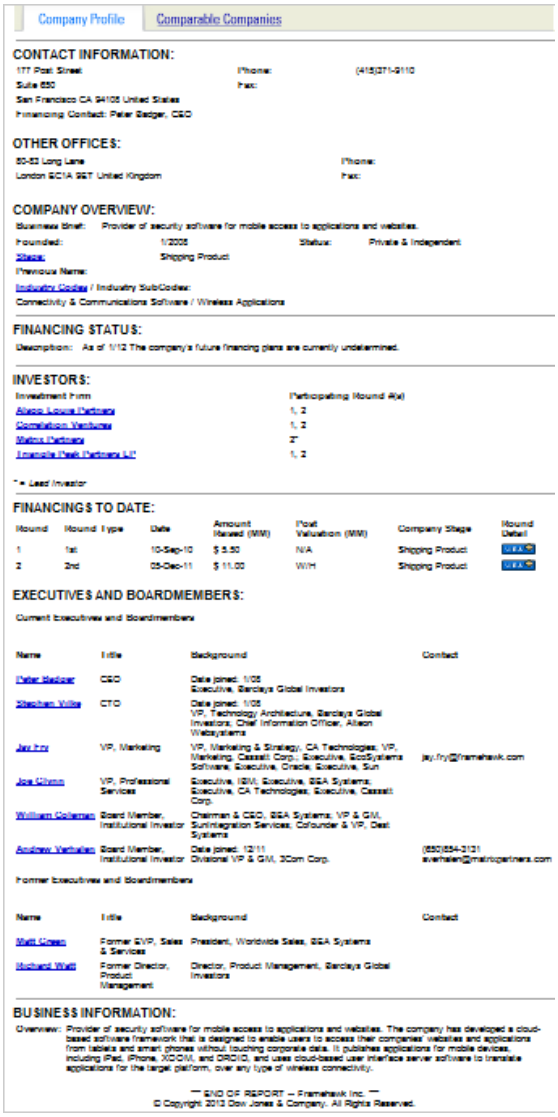
Glossary of Terms.

Term	Definition
Computational analytics	Analytical methods such as classification, optimisation, regression and simulation
Computational tools	Tools that utilise computational methods or techniques
Early-stage investment	Investment classed as pre-seed, seed and startup funding (up to 2m)
Entrepreneur	The owner of a new venture or business
Investment Process	The venture capital investment process, involving deal origination, screening, evaluation, structuring, and post investment activities as the main stages of venture capitalists' decision process.
Investor	An individual (i.e., Investment partner or venture capitalist) or company (i.e., venture capital firm) who invests capital in new ventures
Venture capital	Venture capital is a type of private equity capital typically provided to early-stage, high-potential, growth companies
Venture capital firm	Firms or companies that engage in venture capital type investment activities

Appendix B

Datasets

B.1 VentureSource



Company Profile | Comparable Companies

CONTACT INFORMATION:

177 Post Street
Suite 650
San Francisco CA 94108 United States
Financing Contact: Peter Badger, CEO

Phone: (415)371-2110
Fax:

OTHER OFFICES:

50-53 Long Lane
London EC1A 9ET United Kingdom

Phone:
Fax:

COMPANY OVERVIEW:

Business Brief: Provider of security software for mobile access to applications and websites.
Founded: 1/2005
Status: Private & Independent
Stage: Shipping Product
Previous Name:
Industry Codes / Industry SubCodes:
Connectivity & Communications Software / Wireless Applications

FINANCING STATUS:

Description: As of 1/12 The company's future financing plans are currently undetermined.

INVESTORS:

Investment Firm	Participating Round #(s)
Alloco Equity Partners	1, 2
Constellation Ventures	1, 2
Matrix Partners	2*
Ironside Peak Partners LP	1, 2

* = Lead Investor

FINANCINGS TO DATE:

Round	Round Type	Date	Amount Raised (MM)	Post Valuation (MM)	Company Stage	Round Detail
1	1st	10-Sep-10	$ 5.50	N/A	Shipping Product	VIEW
2	2nd	05-Dec-11	$ 11.00	W/H	Shipping Product	VIEW

EXECUTIVES AND BOARDMEMBERS:

Current Executives and Boardmembers

Name	Title	Background	Contact
Peter Badger	CEO	Date joined: 1/05 Executive, Barclays Global Investors	
Stephen Vitto	CTO	Date joined: 1/05 VP, Technology Architecture, Barclays Global Investors; Chief Information Officer, Allaire Websystems	
Jay Fry	VP, Marketing	VP, Marketing & Strategy, CA Technologies; VP, Marketing, Cassatt Corp.; Executive, EcoSystems Software; Executive, Oracle; Executive, Sun	jay.fry@framehawk.com
Joe Elhmn	VP, Professional Services	Executive, IBM; Executive, BEA Systems; Executive, CA Technologies; Executive, Cassatt Corp.	
William Coleman	Board Member, Institutional Investor	Chairman & CEO, BEA Systems; VP & GM, SunIntegration Services; Cofounder & VP, Best Systems	
Andrew Verhalen	Board Member, Institutional Investor	Date joined: 12/11 Divisional VP & GM, 3Com Corp.	(650)854-3131 averhalen@matrixpartners.com

Former Executives and Boardmembers

Name	Title	Background	Contact
Matt Green	Former EVP, Sales & Services	President, Worldwide Sales, BEA Systems	
Richard Wall	Former Director, Product Management	Director, Product Management, Barclays Global Investors	

BUSINESS INFORMATION:

Overview: Provider of security software for mobile access to applications and websites. The company has developed a cloud-based software framework that is designed to enable users to access their companies' websites and applications from tablets and smart phones without touching corporate data. It publishes applications for mobile devices, including iPad, iPhone, XOOM, and DROID, and uses cloud-based user interface server software to translate applications for the target platform, over any type of wireless connectivity.

— END OF REPORT – Framehawk Inc. —

Figure B.1: VentureSource Screenshot.

B.2 CrunchBase

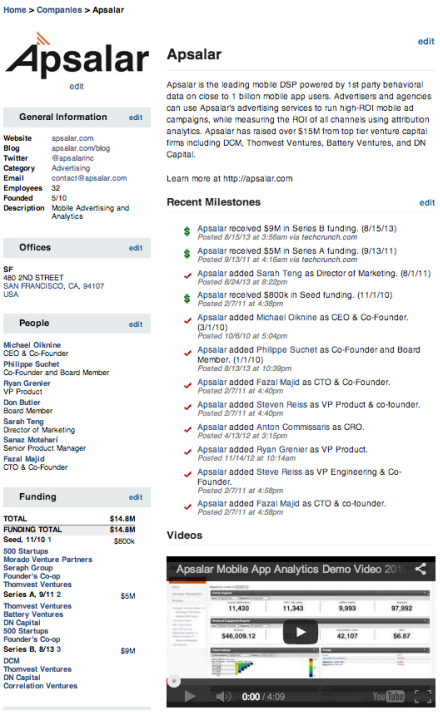


Figure B.2: CrunchBase Screenshot.

B.3 AngelList

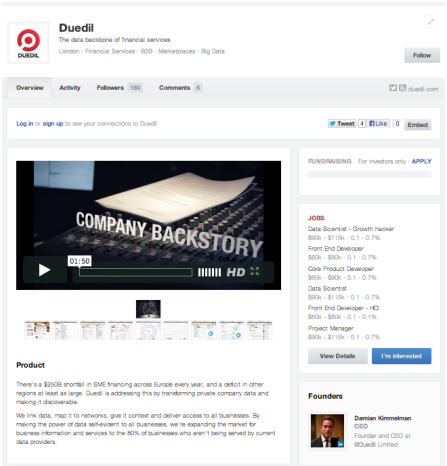


Figure B.3: AngelList Screenshot.

AngelList Dataset

We also have access to a tertiary dataset from AngelList again accessed via publicly available API (Application programming interface) for AngelListAngelList API — https://angel.co/api which provides information on "follows" between investors and companies. AngelList is an active web-based "platform" for angel investors and startups, importantly "following" a startup simply implies interest in a company

not past investment activity. Whilst clearly this is a much weaker indication of relevance compared to actual investment it provides an alternative dataset in order to evaluate matching between companies and investors. In total we have 258,258 startups (i.e., items) and 380,821 users. In regards to relationships we have 1,324,563 relationships.

Table B.1: AngelList Dataset.

Attribute	Description
entity_id	Entity (i.e., company) identity
prin[1-10]	Principal components derived from a singular value decomposition (SVD) of AngelList high level pitch
name	Entity (i.e., company) name
company_url	Entity uniform resource locator (URL) (i.e., website address)
high_level_pitch	Entity high level or short description
tag[1-4]	AngelList market tag classification
user_id	User (i.e., investor or other individual following entity) identity

Appendix C

Additional Experimental Results

C.1 Multi-label Industry Classification

C.1.1 Detailed Results for Binary Classification on VentureSource dataset

Table C.1: Performance of Binary Classification on VentureSource Dataset.

Hierarchy	Class	Accuracy
Group	Information Technology	0.793
Group	Business and Financial Services	0.828
Group	Healthcare	0.942
Group	Consumer Services	0.885
Group	Consumer Goods	0.871
Group	Industrial Goods and Materials	0.877
Group	Energy and Utilities	0.980
Group	To Be Assigned	0.998
Segment	Software	0.794
Segment	Communications and Networking	0.929
Segment	Business Support Services	0.851
Segment	Medical Software and Information Services	0.967
Segment	Consumer Information Services	0.933
Segment	Retailers	0.877
Segment	Electronics and Computer Hardware	0.939
Segment	Travel and Leisure	0.900
Segment	Personal Goods	0.900
Segment	Construction and Civil Engineering	0.976
Segment	Biopharmaceuticals	0.949
Segment	Media and Content	0.940
Segment	Semiconductors	0.951
Segment	Materials and Chemicals	0.908
Segment	Medical Devices and Equipment	0.877
Segment	Wholesale Trade and Shipping	0.997
Segment	Healthcare Services	0.969
Segment	Vehicles and Parts	0.970
Segment	Financial Institutions and Services	0.912
Segment	Renewable Energy	0.983
Segment	Aerospace and Defense	0.969
Segment	Food and Beverage	0.914

Table C.1 – *Continued from previous page*

Hierarchy	Class	Accuracy
Segment	Machinery and Industrial Goods	0.971
Segment	Agriculture and Forestry	0.914
Segment	To Be Assigned	0.998
Segment	Non-Renewable Energy	0.974
Segment	Household and Office Goods	0.983
Segment	Utilities	0.984
Code	Business Applications Software	0.903
Code	Wired Telecommunications Service Providers	0.974
Code	Communications Software	0.913
Code	Design Automation Software	0.974
Code	Wireless Communications Equipment	0.964
Code	Advertising/Marketing	0.953
Code	Clinical Decision Support	0.970
Code	IT Consulting	0.959
Code	Shopping Facilitators	0.952
Code	Specialty Retailers	0.902
Code	Power Supplies	0.980
Code	Automated Manufacturing Equipment	0.962
Code	Incubators/Business Development	0.998
Code	Business to Business Marketplaces	0.982
Code	Travel Arrangement/Tourism	0.940
Code	Computer Systems	0.994
Code	Clothing/Accessories	0.931
Code	General Business Consulting	0.985
Code	Vertical Market Applications Software	0.960
Code	Environmental Engineering/Services	0.975
Code	Drug Development Technologies	0.902
Code	Facilities/Operations Management	0.996
Code	Project/Document Collaboration	0.994
Code	Computer Peripherals	0.996
Code	Internet Service Providers	0.975
Code	Wireless Telecommunications Service Providers	0.970
Code	Business Support Services: Other	0.982
Code	Modems	0.986
Code	Search Portals	0.964
Code	Biotechnology Therapeutics	0.935
Code	General Media/Content	0.975
Code	Wired Communications Equipment	0.951
Code	Application-Specific Integrated Circuits	0.961
Code	Database Software	0.962
Code	Materials and Chemicals: Other	0.921
Code	Consumer Electronics	0.989
Code	Therapeutic Devices Invasive	0.900
Code	Data Storage	0.986
Code	Multimedia/Streaming Software	0.968
Code	Logistics/Delivery Services	0.998
Code	Fiberoptic Equipment	0.973
Code	Network/Systems Management Software	0.951
Code	Recreational/Home Software	0.985
Code	Entertainment	0.965

Table C.1 – *Continued from previous page*

Hierarchy	Class	Accuracy
Code	Memory Systems	0.983
Code	Automated Manufacturing Software	0.994
Code	Procurement/Supply Chain	0.965
Code	Movie/Music Producers and Distributors	0.992
Code	Educational/Training Media and Services	0.935
Code	Elder Care	0.982
Code	Healthcare Administration Software	0.972
Code	Surgical Devices	0.908
Code	Automotive Parts	0.988
Code	Patient Monitoring/Biofeedback	0.961
Code	Email/Messaging	0.988
Code	Customer Relationship Management	0.981
Code	Medical Imaging Software	1.000
Code	Therapeutic Devices Noninvasive	0.888
Code	Educational/Training Software	0.980
Code	Integrated Circuit Production	0.967
Code	Vehicle Parts Retailers/Vehicle Dealers	0.989
Code	Medical Supplies	0.996
Code	Software Development Tools	0.980
Code	Inpatient Facilities	0.982
Code	Medical Devices and Equipment: Other	0.970
Code	Institutional Investment Services	0.960
Code	Filters/Membranes	0.992
Code	Computer Add-On Boards	0.999
Code	Data Management Services	0.985
Code	Software: Other	0.997
Code	Pharmaceuticals	0.902
Code	Personal/Commercial Banking	0.996
Code	Payment/Transactional Processing	0.991
Code	IT Media/Content	0.998
Code	Real Estate	0.955
Code	Human Resources/Recruitment	0.984
Code	Electronics and Computer Hardware: Other	0.998
Code	Lending	0.964
Code	Drug Delivery	0.909
Code	Graphics/Publishing Software	0.994
Code	Diagnostic Equipment Not Imaging	0.892
Code	Systems Software	0.994
Code	Diagnostic Imaging Equipment	0.918
Code	Communications and Networking: Other	0.998
Code	General Purpose Integrated Circuits	0.976
Code	Broadcasting	0.994
Code	Food/Drug Retailers	0.941
Code	Security Services	0.992
Code	Medical Lab Instruments/Test Kits	0.900
Code	Electronic Components/Devices	0.986
Code	Managed Care	0.977
Code	Displays	0.997
Code	Fuel Cells	0.983
Code	Medical/Lab Services	0.980

Table C.1 – *Continued from previous page*

Hierarchy	Class	Accuracy
Code	Commercial Aircraft	0.996
Code	Accounting	0.999
Code	Outpatient Facilities	0.978
Code	Industrial Cleaning Products	0.995
Code	Automobiles	0.996
Code	Media and Content: Other	0.998
Code	Non-Alcoholic Beverages	0.963
Code	General Industrial Goods	0.981
Code	Physician Practice Management	0.976
Code	Commercial Fishing/Aquaculture	0.968
Code	Semiconductors: Other	0.997
Code	Coatings/Adhesives	0.986
Code	Healthcare Services: Other	0.977
Code	To Be Assigned	0.999
Code	Solar Energy	0.990
Code	Home Healthcare Services	0.987
Code	Online Communities	0.961
Code	Restaurants/Food Service	0.958
Code	Conferencing Equipment/Services	0.988
Code	Crop Cultivation/Horticulture	0.987
Code	Agriculture and Forestry: Other	0.928
Code	Biofuels/Biomass	0.971
Code	Exploration Services/Equipment	0.984
Code	Consumer Information Services: Other	0.996
Code	Appliances/Durable Household Goods	0.985
Code	Financial Data/Information	0.991
Code	General Food Products	0.936
Code	Hotels/Gambling	0.989
Code	Wind/Water and Geothermal Energy	0.986
Code	Sports/Leisure Goods	0.947
Code	Insurance	0.989
Code	Clothing/Accessory Retailers	0.934
Code	Specialty Foods	0.935
Code	Personal Care Products	0.976
Code	Retail Investment Services/Brokerages	0.941
Code	Medical Software and Information Services: Other	0.993
Code	Financial Institutions and Services: Other	0.991
Code	Sports/Recreational Services	0.954
Code	Biopharmaceuticals: Other	0.920
Code	Machinery and Industrial Goods: Other	0.989
Code	Communications and Networking: TBA	0.999
Code	Retailers: TBA	1.000
Code	Aircraft Equipment/Parts	0.988
Code	Genetically Modified Agricultural Products	0.987
Code	Gas Distribution	1.000
Code	Legal Counseling	0.988
Code	Health Media/Content	0.990
Code	Photography	0.998
Code	Multiutilities	1.000
Code	Drug Discovery/Bioinformatics Software	0.999

Table C.1 – *Continued from previous page*

Hierarchy	Class	Accuracy
Code	Non-Durable Household Goods	0.968
Code	Agrochemicals	0.983
Code	Vehicles and Parts: Other	1.000
Code	Building Materials/Construction Machinery	0.994
Code	Wholesaling/Distribution Services	0.999
Code	Oil/Gas Exploration and Production	0.975
Code	General Industrial Machinery	0.982
Code	Pet Foods	0.973
Code	Coal	0.997
Code	Electric Utilities	0.985
Code	Biopharmaceuticals: TBA	0.952
Code	Specialty Trade Contractors	0.981
Code	Household Furniture	0.960
Code	Air Freight/Cargo	0.997
Code	Renewable Energy: Other	0.982
Code	Recreational/Sports Vehicles	0.961
Code	Security Products	0.998
Code	Aerospace and Defense: Other	1.000
Code	Ground Freight/Cargo	1.000
Code	Industrial Metals Processing	0.977
Code	Consumer Information Services: TBA	0.999
Code	Plastic Fabrications	0.981
Code	Luxury Goods	0.985
Code	Mixed Retailing	0.976
Code	Basic Chemicals	0.987
Code	Transportation Services	0.995
Code	Personal Services	1.000
Code	Household Goods/Services Retailers	0.999
Code	Dairy Products	0.989
Code	Marine Freight/Cargo	1.000
Code	Home Improvement	0.997
Code	Manufacturing Machinery	1.000
Code	Medical Devices and Equipment: TBA	0.997
Code	Electronics and Computer Hardware: TBA	1.000
Code	Military Vehicles/Aircraft	1.000
Code	Travel and Leisure: Other	0.988
Code	Natural Gas	0.993
Code	Containers/Packaging	0.994
Code	Non-Renewable Energy: TBA	0.992
Code	Personal Goods: Other	0.998
Code	Livestock Farming/Meat Processing	0.996
Code	Industrial Construction	0.999
Code	Textiles	0.978
Code	Food and Beverage: Other	1.000
Code	Utilities: Other	0.994
Code	Software: TBA	0.997
Code	Alcoholic Beverages	0.984
Code	Healthcare Services: TBA	0.995
Code	Building Construction	0.993
Code	Retailers: Other	0.979

Table C.1 – *Continued from previous page*

Hierarchy	Class	Accuracy
Code	Commercial Vehicles	0.998
Code	Office Equipment/Supplies	0.993
Code	Nuclear Energy	0.992
Code	Satellites/Spacecraft	0.999
Code	Semiconductors: TBA	1.000
Code	Household and Office Goods: Other	1.000
Code	Forestry	1.000
Code	Business Support Services: TBA	0.999
Code	Water Utilities	1.000
Code	Non-Renewable Energy: Other	0.998
Code	Architects/Surveyors	1.000
Code	Agricultural Machinery	1.000
Code	Media and Content: TBA	1.000

Appendix D

Extensions

D.1 Discussion of Investment Criteria Literature

Tyebjee and Bruno [TB84] developed a model of the investment activity process, involving deal origination, screening, evaluation, structuring, and post investment activities as the main stages for the venture capitalists actions. These stages coincide with other research into the general stages of investments [FH94]. The factor groups developed of a list of 21 initial factors were taken together to establish five basic characteristics, namely market attractiveness, product differentiation, managerial capabilities, environmental threat resistance, and cash out potential. These factor groups were modeled to either increase expected return (market attractiveness and product differentiation) or decrease the perceived underlying risk of the undertaking (management capabilities and resistance to environmental threats). Interestingly cash out potential was not found to influence risks or returns; therefore it was omitted from the final model. The model was at a later stage of the study validated by introducing several venture capitalists' opinions to check the likelihood of results, most of which were positive with regard to the design. Factors that were said to be underrepresented were the managerial component of the deals, which venture capitalists found to be more important than elaborated in the model. Especially, respondents thought it to have an influence also on the expected return component of the deals. Important to note is the cautious clue towards the heterogeneity of feedback gathered in the evaluation period of the results. Tyebjee and Bruno note that too rigid a specification might improperly equate different venture capitalists' approaches to evaluations. Overall, this study was clear in providing an overview of the important factor groups, identifying market, product, management, cash-out, and survival ratings as the most important.

Macmillan, Siegel, and Subba Narasimha [MSS85] tried to find the most important criteria venture capitalists use to decide on funding new ventures. Six groups of factors could be established that were cited by venture capitalists as important for evaluating new ventures. The factor groups were the entrepreneurs' personality, the entrepreneurs' experience, characteristics of the offering, characteristics of the market, financial considerations, and factors regarding the venture team composition. The finding of

further questioning shows that especially factors regarding the personnel of the new venture are important in making the actual investment decisions. To find underlying patterns of decisions, a cross factor analysis was made and turned out several different classes of risk. These classes were competitive risk (depicting the level of insulation from competition), bail out risk (ease of liquidation of an investment), investment risk (overall probability of failure), management risk (operational capabilities of the management team), implementation risk (actual functioning of offering or business model), and leadership risk (leadership skills of the management team). This data was then used to identify different classes of venture capitalists, representing "purposeful risk managers", carefully evaluating as many risk factors as possible, "determined eclectics", accepting a broad base of investments regardless of some missing criteria, and what they called "parachutists", who invest as long as a liquidity event is relatively certain. The inherent outcome of the study showed a truism in what venture capitalists tend to say about their investments: better to invest in an "A" team with a "B" idea than the other way around. Team composition and management capabilities showed to be the most important factors. One problem with the study is the format of a self-reporting study which does not take actual decisions as base data for the analysis, but rather answers to a questionnaire that was administered by the venture capitalists themselves. In this scenario, answers might be skewed towards what the interviewees might think or perceive to be their reasons, not what are the actual underlying factors. Another flaw is the a priori design of the study which does not take actual investment cases as the basis for the answers, but rather is a hypothetical would be analysis. The take-away for the research at hand are again the factor groups which validate the importance of factors as identified by Tyebjee and Bruno [TB84]. The special focus on the managerial component is a factor that distinguishes this study and supports the criticism of Tyebjee and Bruno, which was said to under represent this factor.

In a follow up to the study, Macmillan, Zemann and Subba Narasimha address some of the mentioned problems by analyzing the most successful and unsuccessful ventures on the basis of a ranking method of several factors [Mac87], thereby eliminating the a- priori design flaws of their first study. The factor groups were venture team, offering, market, and financial forecast characteristics. Also, several performance rankings regarding market and financial factors were measured. The resulting data distinguished several clusters of ventures both in the successful and unsuccessful groups. These clusters were named "well- qualified dropouts", lacking staying power and endurance, "arrow-catchers" for ventures that were first to market, but could not protect their products from competition, and "hapless amateurs", describing ventures that lack in all terms of desirable characteristics. The successful types were called "high-tech sure bets" as the ideal candidates for venture capitalists, "distribution players", usually being involved in more low tech markets, "market makers", who could protect their own part of a market, and "lucky dilettantes", who had some product protection (e.g., patents) that made them successful, but no real success factors otherwise. The researchers found each successful group to match one of the un-

successful ones, only being differentiated by a factor regarding the team composition and capabilities. This is an important finding supporting the previously mentioned focus on team composition, indicating the critical importance of this part of a new venture. In a further analysis, the only two criteria that were found to be relevant for success in the sample under scrutiny were the level of competitive threat and market acceptance. These factors were interestingly not mentioned earlier by the venture capitalists themselves as crucial factors, but turned out to be cut-off points earlier in the due diligence performed on applying firms. Again, self-report issues were a potential problem, but the authors performed a validity check by comparing results to earlier studies, developing results supporting others authors' findings [TB84, MSS85].

Cochrane [Coc00] measures the mean, standard deviation, alpha and beta of venture capital investments. He uses a maximum likelihood estimate that corrects for selection bias on account of firms going public when they have achieved a good return. He states that estimates that do not correct for selection bias are overly optimistic. Cochrane finds returns are very volatile, with a high standard deviation. Furthermore, he manages to substantiate that second, third, and fourth rounds of financing are much less risky. Rounds have progressively lower volatility, lower arithmetic average returns and lower betas.

Riquelme and Watson [RW02] observed how venture capitalists implicit theories about business success are developed into selection criteria to assess the potential viability of new business proposals. They claimed few studies had empirically validated such selection criteria and whether they actually work in relation to future business success. Through reviewing and analysing previous empirical studies into the reasons for small and medium enterprise (SME) success and failure Riquelme and Watson's study supported the selection criteria used.

Ewens [Ewe09] paper characterizes the risk and return of venture capital investments and how they may inform capital allocation using data from VentureSource/VentureOne database. Having formulated a model of venture capital (VC) returns motivated by the entrepreneurial firm life-cycle and the extreme return outcomes of typical venture capital investments, Ewens demonstrates how VC investments offer some risk and return features unavailable in publicly traded equities. Another implication from the study showed that volatility as an estimate of risk, underestimated the frequency and magnitude of large, negative VC returns.

Franke, Gruber, Harhoff, and Henkel [FGHH06] researched biases in venture capital decision making on the basis of similarities of management teams and the deciding investment professionals. They found a rather large coherence between both, a note to be taken on any research towards investment decision factors. Shepherd et al [SZB03], tried to analyze the experience curve of venture capital managers, developing the idea that from a certain point on, more experience in venture capital decision making did not improve performance, but rather diminished it. Zacharakis and Meyer [ZM00], compared actuarial decision models to the decisions made by venture capital professionals and found those models to be of

greater accuracy when judging performance upfront. Zacharakis and Shepherd [ZS01] found out that also more information on a proposal in question did not always lead to better decision making in the long run, but rather made venture capitalists overconfident in their decisions. They argue that the additional information makes decisions more complex, and therefore less reliable outcomes are produced. An important point in all of these papers is the notion of reliability of the decisions and confidence in the outcomes.

D.2 Multi-dimensional Classification

A Business Model Ontology

Another potential alternative would be to base a new classification scheme on existing literature. The work of Osterwalder et al [OP10] in business model ontologies and their business model canvas, a strategic management template for developing new or documenting existing business models provides such an opportunity. The framework is already widely adopted by practitioners and is based upon business model ontology research [Ost04]. Table D.1 depicts the 9-dimensional framework (or building blocks) of the business model canvas and potential class labels for each dimension.

An initial study was undertaken to introduce a novel classification schemes based upon work relating to business models ontologies [Ost04, OP10]. A sample of 264 London technology startups were identified [Due12b, Due12a] and manually classified through an online survey against the 9 dimensions (or "building blocks") of the business model canvas framework by reviewing the individual companies websites. Participants were also ask to define a *market* and *industry* defined as follows [Mul03]:

- *Market* — A market consists of a group of current and/or potential customers having the willingness and ability to buy products goods or services to satisfy a particular class of wants or needs. Thus, markets consist of buyers people or organizations and their needs not products.

- *Industry* — An industry consists of sellers typically organizations which offer products or classes of products that are similar and close substitutes for one another.

In total we received 641 responses with at least 2 responses per company allowing the creation of a training set. We decided if two participants had agreed upon a class for any of the above dimensions it would be suitable to assign that class to that example company.

We then deployed the same supervised classification methodology (see Section 5.3) to classify companies against individual dimensions (e.g., Value Proposition, Customer Relationship) of this novel multi-dimensional classification scheme.

Whilst we had limited success (see Appendix C) in accurately auto classifying companies based on cross-validation and the same experimental settings and measures (see Section 5.5) we believe a much larger training set is necessary to further this line of enquiry. Furthermore, arguably only a subset of these

Table D.1: Business Model Ontology [Ost04, OP10].

Dimension	Description	Classes
Customer Segments	The Customer Segments building block defines the different groups of people or organisations an enterprise aims to reach and serve	Mass market Niche market Segmented Diversified Multi-sided platforms
Value Proposition	The Value Propositions building block describes the bundle of products and services that create value for a specific Customer Segment	Newness Performance Customization ”Getting the job done” Design Price Cost reduction Risk reduction Accessibility Convenience/usability
Channels	The Channels building block describes how a company communicates with and reaches its Customer Segments to deliver a Value Proposition	Sales force Web sales Own stores Partner stores Wholesaler
Customer Relationships	The Customer Relationships building block describes the types of relationships a company establishes with specific Customer Segments	Personal assistance Dedicated personal assistance Self-service Automated services Communities Co-creation
Revenue Streams	The Revenue Streams building block represents the cash a company generates from each Customer Segment	Asset sale Usage fee Subscription fees Lending/Renting/Leasing Licensing Brokerage fees Advertising
Key Resources	The Key Resources building block describes the most important assets required to make a business model work	Physical Intellectual Human Financial
Key Activities	The Key Activities building block describes the most important things a company must do to make its business model work	Production Problem solving Platform/network
Key Partnerships	The Key Partnerships building block describes the network of suppliers and partners that make the business model work	Optimization and economy of scale Reduction of risk and uncertainty Acquisition of particular resources and activities
Cost Structure	The Cost Structure building block describes all costs incurred to operate a business model	Cost-driven Value-driven

dimensions can be ascertained from inspection and textual analysis of a company's website. For example, building blocks such as Value Proposition, Channels and Customer Relationship are likely more identified and visible on a company's website then for example Cost Structure. This became evident from the responses provided in our original survey. Finally, the free text responses also highlighted the confusion and ambiguity associated with the terms "industry" and "market". For example, for the company 7digital self-described as "a privately held digital music platform, providing access to over 23 million legal, high quality tracks"[1] was associated with 5 different proposed markets ("Consumers", "Mobile", "Music listeners", "Music lovers', "open digital content platform targeting people who are looking for high quality digital products") and 5 different proposed industries ("Music/Recording", "mobile Apps", "Online music stores", "Music selling platform", "IT").

D.3 User Study

In terms of extensions, as previously mentioned we designed two distinct user studies in collaboration with industry partner Correlation Ventures, a venture capital firm, in order to test our proposed use cases in venture finance.

- **Peer Identification and Company Similarity Measures** — in order to analyse competitors and compare potential investment opportunities (covered in Chapter 6).
- **Top-N Investment Opportunity Recommendation** — for the purpose of finding co-investment partners and identifying the most suitable investors for a company and vice versa based upon past investment history (covered in Chapter 7.

Our second user study was devised to test our recommendation models with a small number of investment partners at Correlation Ventures. Again, we plan to make use of our access to the VentureSource dataset and test three different recommendation techniques which all rely on information about an investor's (i.e., investment partner) prior investments. Therefore our compared techniques for recommending investment opportunities would likely be:

1. A content-based approach using kNN over industry Code vector
2. A collaborative filtering approach CF using kNN
3. A hybrid approach combing CF and Multi Code both using kNN

Using the above techniques we recommend the top 20 investment opportunities with each technique for each investment partner. Each investment partner is then provided as worksheet with the recommended companies and and a field for providing a relevance score between 1 (low) and 5 (high).

[1] 7digital — `http://www.7digital.com/`

Respondents are required to review the description of each company and select a relevance score between $\star$ (low) and $\star\star\star\star\star$ (high) based on to what extent they agree with the statement *"This company is relevant as a potential investment opportunity"* based on the following scale:

$\star$ — Strongly disagree

$\star\star$ — Disagree

$\star\star\star$ — Neither agree nor disagree

$\star\star\star\star$ — Agree

$\star\star\star\star\star$ — Strongly agree

This user study is currently incomplete and would be beneficial in further validating (or invalidating) the use and efficacy of such recommendation techniques in regards to Top-N investment opportunity recommendation.

Appendix E

Algorithms

E.1 Naive Bayes

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes theorem with the naive assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through x_n, Bayes theorem states the following relationship [Sl10b]:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive independence assumption that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$.

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They requires a small amount of training data to estimate the necessary parameters. (For theoretical reasons why naive Bayes works well, and on which types of data it does, see [Zha04])

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

On the flip side, although naive Bayes is known as a decent classifier, it is known to be a bad estimator, so the probability outputs are not to be taken too seriously.

E.2 c4.5

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [Sl10c].

Decision trees learn from data to approximate a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

ID3 (Iterative Dichotomiser 3) was developed in 1986 by Ross Quinlan. The algorithm creates a multiway tree, finding for each node (i.e. in a greedy manner) the categorical feature that will yield the largest information gain for categorical targets. Trees are grown to their maximum size and then a pruning step is usually applied to improve the ability of the tree to generalise to unseen data.

C4.5 is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. The accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rule's precondition if the accuracy of the rule improves without it.

Given training vectors $x_i \in R^n, i = 1, ..., l$ and a label vector $y \in R^l$, a decision tree recursively partitions the space such that the samples with the same labels are grouped together. Let the data at node m be represented by Q. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ subsets

$$Q_{left}(\theta) = (x, y)|x_j <= t_m$$

$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$$

The impurity at m is computed using an impurity function $H()$, the choice of which depends on the task being solved (classification or regression)

$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$

Select the parameters that minimises the impurity

$$\theta^* = \operatorname{argmin}_\theta G(Q, \theta)$$

Recurse for subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until the maximum allowable depth is reached, $N_m < \min_{samples}$ or $N_m = 1$.

If a target is a classification outcome taking on values $0, 1, ..., K - 1$, for node m, representing a region R_m with N_m observations, let

$$p_{mk} = 1/N_m \sum_{x_i \in R_m} I(y_i = k)$$

be the proportion of class k observations in node m

Common measures of impurity are Gini

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

Cross-Entropy

$$H(X_m) = \sum_k p_{mk} \log(p_{mk})$$

and Misclassification

$$H(X_m) = 1 - \max(p_{mk})$$

E.3 Random Forests

The goal of ensemble methods is to combine the predictions of several models built with a given learning algorithm in order to improve generalisability / robustness over a single model [Sl10d].

Two families of ensemble methods are usually distinguished:

- In averaging methods, the driving principle is to build several models independently and then to average their predictions. On average, the combined model is usually better than any of the single model because its variance is reduced.
  - Examples: Bagging methods, Forests of randomised trees...
- By contrast, in boosting methods, models are built sequentially and one tries to reduce the bias of the combined model. The motivation is to combine several weak models to produce a powerful ensemble.
  - Examples: AdaBoost, Gradient Tree Boosting, ...

There are various algorithms based on randomised decision trees. This means a diverse set of classifiers is created by introducing randomness in the classifier construction. The prediction of the ensemble is given as the averaged prediction of the individual classifiers.

As other classifiers, forest classifiers have to be fitted with two arrays: an array X of size n (samples, features) holding the training samples, and an array Y holding the target values (class labels) for the training samples.

In random forests each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to

averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.

E.4 Support Vector Machines

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection [Sl10a].

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, the method is likely to give poor performances.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

Given training vectors $x_i \in R^p, i = 1, ..., n$, in two classes, and a vector $y \in R^n$ such that $y_i \in \{1, -1\}$, SVC solves the following primal problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1,n} \zeta_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0, i = 1, ..., n$$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$

$$\text{subject to } y^T \alpha = 0$$

$$0 \leq \alpha_i \leq C, i = 1, ..., l$$

where e is the vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv K(x_i, x_j)$ and $\phi(x_i)^T \phi(x)$ is the kernel. Here training vectors are mapped into a higher (maybe infinite) dimensional space by the function ϕ.

The decision function is:

$$\text{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho)$$

Note: While SVM models derived from libsvm and liblinear use C as regularisation parameter, most other estimators use alpha. The relation between both is $C = \frac{n\_samples}{alpha}$.

Bibliography

[AA12] Arthur Absalom and Geoffrey Absalom. Durham Zoo : prior art and solution search. Technical report, 2012.

[ABSW11] OT Alexy, JH Block, Philipp Sandner, and ALJ Ter Wal. Social capital of venture capitalists and start-up funding. *Small Business Economics*, 2011.

[AfSG11] Nino Antulov-fantulin, Tomislav Smuc, and Dragan Gamberger. Constructing recommender systems workflow templates in RapidMiner. 2011.

[AH00] Lada A Adamic and Bernardo A Huberman. Power-law distribution of the world wide web. *Science*, 2000.

[Alb04] Russ Albright. Taming Text with the SVD. Technical report, SAS Institute Inc., 2004.

[And11a] Marc Andreessen. Marc Andreessen: Structural Changes in Venture Capital, 2011.

[And11b] Marc Andreessen. Why Software Is Eating the World, 2011.

[Ang13a] AngelList. AngelList, 2013.

[Ang13b] AngelList. AngelList, 2013.

[Ant12] Scott Anthony. Is Venture Capital Broken?, 2012.

[Aud95] DB Audretsch. Innovation, growth and survival. *International journal of industrial organization*, 13:441–457, 1995.

[Ban91] Graham Bannock. *Venture capital and the equity gap*. Graham Bannock & Partners, 1991.

[Bar94] Christopher B Barry. New Directions in Research on Venture Capital Finance. *Financial Management*, 23(3):3–15, 1994.

[BEP09] M Bishop, S Engle, and S Peisert. WE HAVE MET THE ENEMY . . . AND HE IS US . (May), 2009.

[BFG11] Robin Burke, Alexander Felfernig, and MH Göker. Recommender systems: An overview. *AI Magazine*, pages 13–18, 2011.

[BL10] Paul Belleflamme and Thomas Lambert. Crowdfunding : An Industrial Organization Perspective. *Business*, 2010.

[Bla09] Steven Gary Blank. Customer Development Manifesto: Market Type, 2009.

[Bla13] Steve Blank. Its Time to Play Moneyball: The Investment Readiness Level — Steve Blank on WordPress.com, 2013.

[BLO] Sanjeev Bhojraj, Charles M. C. Lee, and Derek K. Oler. *Journal of Accounting Research*.

[BS11] Tarun Bhaskar and Gopi Subramanian. Loan recommender system for microfinance loans: Increasing efficiency to assist growth. *Journal of Financial Services Marketing*, 15(4), 2011.

[BVC11] BVCA. BVCA : What are private equity and venture capital?, 2011.

[BW97] G. Boocock and M. Woods. The Evaluation Criteria used by Venture Capitalists: Evidence from a UK Venture Fund. *International Small Business Journal*, 16(1):36–57, October 1997.

[BYRN99] R Baeza-Yates and B Ribeiro-Neto. *Modern information retrieval*. 1999.

[Cai11] Claire Cain Miller. Google Spending Millions to Find the Next Google, 2011.

[CKL05a] Bart Clarysse, Mirjam Knockaert, and Andy Lockett. How do Early Stage High Technology Investors Select Their Investments? *Venture Capital*, 2005.

[CKL05b] Bart Clarysse, Mirjam Knockaert, and Andy Lockett. How do early stage high technology investors select their investments. Technical report, Ghent University, Faculty of Economics and Business Administration, 2005.

[CL11] Chih-Chung Chang and Chih-Jen Lin. {LIBSVM}: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1—-27:27, 2011.

[Coc00] John Cochrane. The Risk and Return of Venture Capital. 2000.

[Cor13] Correlation Ventures. Correlation Ventures - Our Approach, 2013.

[Cru13a] CrunchBase. CrunchBase - Welcome to the CrunchBase Developer Portal, 2013.

[Cru13b] CrunchBase. CrunchBase, The Free Tech Company Database, 2013.

[CY13] Alistair Croll and Benjamin Yoskovitz. *Lean Analytics*. O'Reilly Media, 2013.

[CZL12] Tianqi Chen, W Zhang, and Qiuxia Lu. SVDFeature: A Toolkit for Feature-based Collaborative Filtering. *JMLR*, 2012.

[Daf07] Richard L Daft. Understanding the theory and design of organizations. 2007.

[Dew76] Melvil Dewey. *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library [Dewey Decimal Classification]*. 1876.

[DK04] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004.

[Due12a] DueDil. Duedil's List of London Startups, 2012.

[Due12b] DueDil. The Real London Tech Startups, 2012.

[Ehr11] Roger Ehrenberg. Information Arbitrage - Quantitative methods and seed stage investing, 2011.

[Ern13] Ernst & Young. Turning the corner. Global venture capital insights and trends 2013. Technical report, January 2013.

[Eur11] European Venture Capital Association (EVCA). EVCA Yearbook 2011. 2011.

[Ewe09] Michael Ewens. A New Model of Venture Capital Risk and Return. Technical Report June, Working paper, UCLA., 2009.

[Faw06] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[FGHH06] Nikolaus Franke, Marc Gruber, Dietmar Harhoff, and Joachim Henkel. What you are is what you likesimilarity biases in venture capitalists' evaluations of start-up teams. *Journal of Business Venturing*, 21(6):802–826, November 2006.

[FH94] Vance H. Fried and Robert D. Hisrich. Toward a Model of Venture Capital Investment Decision Making. *Financial Management*, 23(3):28–37, 1994.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. 2008.

[Fro11] Dan Frommer. Moneyball for tech startups, 2011.

[Ger13] Tomio Geron. Real-Time Venture Capital Investing: Chi-Hua Chien's Data-Based Approach - Forbes, 2013.

[GMPR07] David Gill, Tim Minshall, Craig Pickering, and Martin Rigby. Funding Technology - Britiain Forty Years On. *Technology*, (January), 2007.

[GR11] Z Gantner and Steffen Rendle. MyMediaLite: A free recommender system library. 2011.

[Gra05] Paul Graham. How to Fund a Startup, 2005.

[Gra13] Paul Graham. Startup Investing Trends, 2013.

[Gro13] GroupLens Research. MovieLens Data Sets, 2013.

[HKTR04] Jonathan L. Herlocker, Joseph a. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, January 2004.

[HVS+12] Mario Hernandez, Jordi Vitria, Joao Miguel Sanches, Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012.

[Ins13] CB Insights. Venture Capital Database - CB Insights, 2013.

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, pages 42–49, 2009.

[Kni21] Frank Hyneman Knight. *Risk, Uncertainty and Profit*. 1921.

[Kor] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*.

[Kor08] Yehuda Koren. Factorization meets the neighborhood. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 426, New York, New York, USA, August 2008. ACM Press.

[KSS09] Steven N Kaplan, Berk A Sensoy, and Per Strömberg. Should investors bet on the jockey or the horse? evidence from the evolution of firms from early business plans to public companies. *The Journal of Finance*, 64(1):75–115, 2009.

[Law13] Ryan Lawler. VC Firm E.ventures Makes Its Internal Tracking Tool, The Daily Gieselmann, Available To All — TechCrunch, 2013.

[LLP01] Choonwoo Lee, Kyungmook Lee, and Johannes M Pennings. Internal capabilities, external networks, and performance: a study on technology-based ventures. *Strategic management journal*, 22(6-7):615–640, 2001.

[LS10] Thomas Lambert and Armin Schwienbacher. An Empirical Analysis of Crowdfunding. *Analysis*, pages 1–23, 2010.

[Mac87] I A N C Macmillan. Criteria distinguishing successful from unsuccessful Ventures in the Venture Screening Process. *Seven*, 137:123–137, 1987.

[MAFv12] Matej Mihelčić, Nino Antulov-Fantulin, and Tomislav Šmuc. Rapid Miner Recommender Extension - user guide, 2012.

[Man02] S Manigart. Determinants of required return in venture capital investments: a five-country study. *Journal of Business Venturing*, 17(4):291–312, July 2002.

[Mat13] Mattermark. Mattermark - Quantifying Private Company Growth for Startup Investors, 2013.

[Met07] Andrew Metrick. *Venture Capital and the Finance of Innovation*. 2007.

[MHD+12] Max Marmer, B Herrmann, E Dogrultan, R Berman, C Eesley, and S Blank. Startup genome report extra premature scaling. *Startup Genome*, 2012.

[MHSM02] Brent Mainprize, Kevin Hindle, Brock Smith, and Ron Mitchell. TOWARD THE STANDARDIZATION OF VENTURE CAPITAL INVESTMENT EVALUATION : DECISION CRITERIA FOR RATING INVESTEE BUSINESS PLANS. (Pareto 1896):1–11, 2002.

[Mon03] Miquel Montaner. A Taxonomy of Recommender Agents on the Internet. pages 285–330, 2003.

[Moo99] Geoffrey A Moore. *Inside the tornado: marketing strategies from Silicon Valley's cutting edge*. HarperPerennial, 1999.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. July 2008.

[MSS85] I A N C Macmillan, Robin Siegel, and P N Subba Narasimha. CRITERIA USED BY VENTURE CAPITALISTS TO EVALUATE NEW VENTURE PROPOSALS. 128:119–128, 1985.

[Mul03] John Mullins. *The New Business Road Test: What Entrepeneurs and Executives Should Do Before Writing a Business Plan*. Financial Times/ Prentice Hall, 2003.

[Nat09] National Institute of Standards and Technology. Common Evaluation Measures. In *The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings*, 2009.

[NMC+09] Paul Nightingale, Gordon Murray, Marc Cowling, Baden-Fuller Charles, Coloin Mason, Josh Siepel, Mike Hopkins, and Charles Dannreuther. From funding gaps to thin markets - UK Government support for early-stage venture capital. Technical Report September, 2009.

[NR13] Max Nathan and Anna Rosso. MEASURING THE UK'S DIGITAL ECONOMY WITH BIG DATA. Technical report, The National Institute of Economic and Social Research (NIESR), 2013.

[Off07] Office for National Statistics (ONS). UK Standard Industrial Classification 2007 (UK SIC 2007), 2007.

[Ogu13] Matt Oguz. The Moneyball strategy is the future for venture capital firms, 2013.

[OP10] Alexander Osterwalder and Yves Pigneur. *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. John Wiley & Sons, 2010.

[Ost04] Alexander Osterwalder. *THE BUSINESS MODEL ONTOLOGY A PROPOSITION IN A DESIGN SCIENCE APPROACH*. PhD thesis, lUniversité de Lausanne, 2004.

[oT07] Lord Sainsbury of Turville. The Race to the Top. *October*, (October), 2007.

[PB07] Michael J Pazzani and Daniel Billsus. Content-Based Recommendation Systems. pages 325–341, 2007.

[PK89] Bruce D. Phillips and Bruce a. Kirchhoff. Formation, growth and survival; Small firm dynamics in the U.S. Economy. *Small Business Economics*, 1(1):65–74, 1989.

[Rav12] Naval Ravikant. The House Financial Services Committee The House Oversight And Government Reform Committee Joint Subcommittee Hearing on JOBS Act Implementation, 2012.

[RB57] Everett M Rogers and George M Beal. Importance of personal influence in the adoption of technological change, the. *Soc. F.*, 36:329, 1957.

[RFGST09] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, 2009.

[RGFST11] Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme. Fast context-aware recommendations with factorization machines. *SIGIR*, 2011.

[Rig11] Sarah Rigos. The UK equity gap. (November), 2011.

[RR92] Hernan Riquelme and Tudor Rickards. HYBRID CONJOINT ANALYSIS : AN ESTIMATION PROBE IN NEW VENTURE DECISIONS. *Analysis*, pages 505–518, 1992.

[RW02] H. Riquelme and J. Watson. Do Venture Capitalists' Implicit Theories on New Business Success/Failure have Empirical Validity? *International Small Business Journal*, 20(4):395–420, November 2002.

[SC02] Scott Shane and Daniel Cable. Network ties, reputation, and the financing of new ventures. *Management Science*, 48(3):364–381, 2002.

[Sch11] Erick Schonfeld. The Top 10 VC Firms, According To InvestorRank — TechCrunch, 2011.

[See] Seedcamp. Seedcamp Results: already best in class in Europe, now also decidedly world-class — Seedcamp.

[SHH99] Toby E Stuart, Ha Hoang, and Ralph C Hybels. Interorganizational endorsements and the performance of entrepreneurial ventures. *Administrative science quarterly*, 44(2):315–349, 1999.

[SK09] Xiaoyuan Su and Taghi M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 2009(Section 3):1–19, 2009.

[Sl10a] Scikit-learn. 1.2. Support Vector Machines scikit-learn 0.14 documentation, 2010.

[Sl10b] Scikit-learn. 1.7. Naive Bayes scikit-learn 0.14 documentation, 2010.

[Sl10c] Scikit-learn. 1.8. Decision Trees scikit-learn 0.14 documentation, 2010.

[Sl10d] Scikit-learn. 1.9. Ensemble methods scikit-learn 0.14 documentation, 2010.

[SM86] Gerard Salton and Michael J. McGill. Introduction to Modern Information Retrieval. October 1986.

[SM02] SD Sarasvathy and Anil Menon. Failing firms and successful entrepreneurs: Serial entrepreneurship as a simple machine. *Unpublished manuscript*, (412), 2002.

[Sor03] Morten Sorensen. How Smart is Smart Money ? An Empirical Two-Sided Matching Model of Venture Capital. *Economic Policy*, 2003.

[SPUP02] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. *SIGIR*, 2002.

[SS01] Olav Sorenson and Toby E Stuart. Syndication networks and the spatial distribution of venture capital investments1. *American Journal of Sociology*, 106(6):1546–1588, 2001.

[Sta12] Stanford University Office of Technology Licensing. Start-up Guide. 2012.

[STL11] G Schröder, Maik Thiele, and Wolfgang Lehner. Setting Goals and Choosing Metrics for Recommender System Evaluations. In *UCERSTI 2 - ACM Conference on Recommender Systems (RECSYS 2011)*, pages 78–85, 2011.

[SZB03] Dean A. Shepherd, Andrew Zacharakis, and Robert A. Baron. VCs' decision processes. *Journal of Business Venturing*, 18(3):381–401, May 2003.

[TB84] Tyzoon T Tyebjee and Albert V Bruno. A Model of Venture Capitalist Investment Activity. *Management Science*, 30(9):1051–1066, 1984.

[TS06] Jeffry Timmons and Stephen Spinelli. *New Venture Creation: Entrepreneurship for the 21st Century with Online Learning Center access card*. McGraw-Hill/Irwin, 2006.

[Win13] Greg Winterton. More VCs bet on algorithms to mine for deals, 2013.

[WVSB11] Willem Waegeman, Jan Verwaeren, Bram Slabbinck, and Bernard De Baets. Supervised learning algorithms for multi-class classification problems with partial class memberships. *Fuzzy Sets and Systems*, 184(1):106–125, 2011.

[Y C12] Y Combinator. New: Apply to Y Combinator without an Idea, 2012.

[YRAS01] Helena Yli-Renko, Erkko Autio, and Harry J Sapienza. Social capital, knowledge acquisition, and knowledge exploitation in young technology-based firms. *Strategic management journal*, 22(6-7):587–613, 2001.

[YTN10] Chaiyakorn Yingsaeree, Philip Treleaven, and Giuseppe Nuti. COMPUTATIONAL FINANCE. *IEEE Computer*, 2010.

[ZCWY13] W. Zhang, T. Chen, J. Wang, and Yong Yu. Optimizing top-n collaborative filtering via dynamic negative item sampling. In *SIGIR*, 2013.

[Zha04] Harry Zhang. The optimality of naive Bayes. *AA*, 2004.

[Zho12] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall, 2012.

[ZM98] Andrew L Zacharakis and G Dale Meyer. A LACK OF INSIGHT : DO VENTURE CAPITALISTS REALLY UNDERSTAND THEIR OWN DECISION PROCESS ? *Journal of Business Venturing*, 9026(97):57–76, 1998.

[ZM00] Andrew L Zacharakis and G Dale Meyer. The Potential of Actuarial Decision Models: Can they improve the venture capital investment decision? *Journal of Business Venturing*, 9026(98):323–346, 2000.

[ZS01] Andrew L Zacharakis and Dean A Shepherd. The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, 16(4):311–332, July 2001.