Working Paper Series

Nesta Working Paper No. 16/02

Quantifying entrepreneurship using domain name registration data: Methods and Applications for Oxford, UK and Cambridge, UK

Abhishek Nagaraj
Sibo Wang

Quantifying Entrepreneurship Using Domain Name Registration Data: Methods and Applications for Oxford, UK and Cambridge, UK

Abhishek Nagaraj
UC Berkeley / NBER

Sibo Wang
Massachusetts Institute of Technology

Nesta Working Paper
16/02 September, 2016

www.nesta.org.uk/wp16-02

Abstract

What is the rate of entrepreneurship around the world? How does it differ by country, by region and by sector? Without a comprehensive, centralized and global registry for new firm founding this question is difficult to answer. In this paper, we make progress on this issue by using a central insight \u2013 most growth-oriented businesses around the world begin by registering a domain name and launching a website. We propose a novel method based on leveraging the domain name system registration system central to the operation of the internet to provide a measure of regional entrepreneurship. Using machine learning and text-analysis techniques, we show how domain name registration data could be used to count new firm formation at an extremely granular level both over time and at the micro-geographic level. We provide a sample application of these ideas to measure new firm formation in Oxford, UK and Cambridge, UK. This work highlights the promise of using domain name registration data to measure entrepreneurial activity across many regions of the world, especially when systematic business registration data are unavailable. This work complements and improves upon a number of existing such efforts including the Global Entrepreneurship Monitor, World Bank New Business density database and the European Digital City Index.

JEL Classification: R12, O12, O3

Keywords: Entrepreneurial ecosystem measurement, regional ecosystems, domain name systems, entrepreneurship in the UK

We are thankful to Christopher Haley for providing insightful feedback and encouragement for this project and to NESTA for their financial support. We also thank Ed Freyfogle of OpenCageData for his help with the geocoding algorithm and Elina Gaillard, Valerie Mocker and Liam Collins for their comments and feedback.

Correspondence: Abhishek Nagaraj, University of California at Berkeley, Haas School of Business, 565 Student Services Building, Berkeley, CA 94720-1900, nagaraj@berkeley.edu

Contents

1. Introduction

Entrepreneurship is the central process through which economic growth and performance is fostered in a regional economy. Both economists and policymakers have recognized the importance of entrepreneurship to economic growth, and have debated a variety of different policies and measures to encourage entrepreneurship at the regional level. However, the evaluation of the effectiveness of these policies has been bedeviled by the inability to measure growth-oriented entrepreneurship. In addition to their importance to measuring the effectiveness of policies meant to spur entrepreneurship, such measures are also necessary to understand the geographic and temporal variation in entrepreneurial activity in any given region.

Ultimately, a lack of timely and granular data to measure entrepreneurship means that it is difficult to measure attempts to start growth-oriented businesses, and then evaluate their likelihood of success through measures like their ability to raise venture capital or IPO and their revenues or sales.

A number of efforts have been made by private enterprises and think tanks to develop regional indices of startup activity including the Global Entrepreneurship Monitor (GEM) (Reynolds et al. 2000), the World Bank New Business density database (for eg: Klapper et al, 2010), the NESTA European Digital City Index

(EDCi) (see Haley, 2015) and the OpenCorporates Open Company Data Index (OCDI). Recent academic work (Guzman and Stern 2015) has also started to make progress on this question by using business registration data. In this work, the authors leverage complete business registration lists and then match these data with a number of related covariates to predict growth-oriented entrepreneurship.

While this method is quite effective, it relies crucially on the availability of business registration data. While such data is readily available in some states in the US, in many other states such data is either simply not collected, is highly fragmented, or is unavailable to researchers.
Similarly, the indices developed for Europe such as the OCDI also suffer from a lack of consistent data and poor coverage in certain geographies. Beyond the US and Europe, and especially in the developing world, this problem is even more acute. Therefore such methods cannot be applied to many regions around the world where there is a sense that entrepreneurship is flowering and policy-makers are quite interested in understanding growth-oriented entrepreneurship.

This paper aims to develop a methodology that aims to address this gap. The method we propose promises to be able to estimate entrepreneurship activity at a global scale. Our aim is to develop a single and centralized technique and will work across geographies – including the developing and the developed world. Having said that the methods we propose will work better in certain contexts as compared to others, and need to be fine-tuned to local specifications. However, a majority of the machinery we develop could be applied across geographies with only minor modifications.

This methodology is built on the central insight that most growth-oriented businesses begin not with company registration, but with “domain-name” registration. Entrepreneurs often establish a website for their business in order to start promoting it, and to get “business” email addresses, and we show that this activity can be a useful predictor of entrepreneurship activity. In this project, we leverage this central insight to provide a new way to measure global entrepreneurship. As we will show, domain name registration data could be extremely insightful and have many appealing properties. They are accessible through central repositories because of internet governance regulation, they are global (all countries have to follow the same domain name registration rules), they are available historically and they link the domain name of the business to variables such as the name of the registrants (the founders), the location of the registrants and other useful information. Perhaps their greatest difference from business registration data is the ability to match the domain name to the text of the web-page that the domain name hosts. This text can be extremely rich and provides a direct indicator of the entrepreneurial intent of the company, including the sector of the business and the growth-orientation of the company.

In this paper, we will illustrate how internet registration records could be used to predict growth-oriented entrepreneurship. We will use domain name records from Oxford and Cambridge in the UK to illustrate our methodology. These methods can then be used to develop easily digestible metrics and dashboards that can then be used to drive debate around innovation policy and increase the general public's interest in this topic. For example, this technique could be used to develop a timely "entrepreneurship index", or sector-specific or region-specific indices (similar to indices such at the EDCi) that would help policymakers, and the general public understand the emerging innovation landscape on the internet. Finally, this database would also be made available so that other researchers and policymakers could use it for follow-on research, and to test the effectiveness of regional entrepreneurship policies.

2. Methodology

2.1 Overview

The methodology for this project relies on the technical structure of the internet to build a database of internet-enabled innovation. An important insight is that on the internet, many potentially revolutionary idea begins with a simple action – “domain name registration” – for example, Google began with a registration for the domain “google.com” which domain name records indicate was created on 15th September 1997, whilst Google as a company was not incorporated until September 1998. While this information is currently accessible in a limited manner (i.e. there are limits on how many records can be obtained at one time), technically it is public information that could be organized in a central repository. Further there are a number of data aggregators and providers who have already aggregated this information and parsed it in a manner that is ready for statistical analysis. At the heart of these data is the “whois” protocol.

The whois protocol which was specified in RFC 3912 (a publication from the Internet Engineering Task Force which suggests Internet standards), and states its function:

> *WHOIS is a TCP-based transaction-oriented query/response protocol that is widely used to provide information services to Internet users. While originally used to provide "white pages"*

> *services and information about registered domain names, current deployments cover a much broader range of information services. The protocol delivers its content in a human-readable format.*

Each domain name corresponds to a WHOIS record which provides a host of useful information relevant to measuring entrepreneurship. For example, an extract of the Google's current WHOIS record shown below includes the following details:

> *GOOGLE WHOIS RECORD*
> *Updated Date:*
> *20140519T04:00:170700 Creation*
> *Date: 19970915T00:00:000700*
> *Registrant Organization: Google Inc.*
> *Registrant Street: Please contact, 1600 Amphitheatre*
> *Parkway Registrant City: Mountain View*
> *Registrant State/Province:*
> *CA Registrant Postal Code:*
> *94043 Registrant Country:*
> *US*
> *Registrant Phone: +1.6502530000*

Notice how this record contains the time at which the google.com domain was registered, and the possible location of the firm that was responsible for the registration. Further, using web-scraping and other techniques, further information can be obtained about the domain google.com – which would indicate that the domain name was an internet service in the area of search engines. As is evident, these records include information on the firms and individuals registered against a given domain, their street address and other contact details. These “whois” records can, therefore, be used to track who owns many websites and the exact city of operation. Combined, these records make it easy to access data on when certain domain names were created as well as the name and physical address of the firm or individual that created them, thereby linking a possibly innovative new internet service to a physical location and to individual firms.

We worked with one aggregator of these data, WHOISXMLAPI which provided us with whois records that were registered in the cities of Cambridge and Oxford in the UK. Whois data are human-readable (like the example above), but they need to be parsed, so that each individual field can be separately parsed by a statistical program. WHOISXMLAPI has written algorithms and programs that provide already parsed data to the end user. These data include the following broad fields: administrativeContact, billingContact, registrant,technicalContact, zoneContact among others. For each of these “contact” fields, the following information is provided: name, organization, street1, street2, city, state, postalCode. An example record for the domain “jombay.com”, an innovative HR technology company based in Pune, India is given below:

"name": "Suruchi Wagh",

"organization": "Next Leap Career Solutions Pvt. Ltd",

"street1": "301, Saikar Paradise\nNear Sapling School,
Baner", "city": "Pune",
"state":
"Maharashtra",
"postalCode":
"410045", "country":
"INDIA",
"email": "suruchi@jombay.com",
"telephone": "91919987979029",
"createdDate":
"2012-09-29T06:16:40Z",

This whois record contains significant information on the domain jombay.com. It tells you Suruchi Wagh probably registered (and likely founded) a company called Jombay when she was based in Pune, India in September 2012. Collecting and harvesting a database of all such domains registered in Oxford and Cambridge is the first step of the process for us.

However, simply compiling a complete list of every single domain registered in Oxford and Cambridge would be inadequate because it would include many domains and websites that are not of interest to measuring entrepreneurial activity. The domains include personal webpages, websites for nonprofits, agencies and media and news websites. In order to address this issue and to filter the list of domains, we used additional information about the domain name from different sources. First, we scraped the public websites for these domain names to get a sense of the webpages that were being served under these URLs. We also used the web archive (archive.org) to scrape historical web page data for many domain names. Finally, domain names are being tracked by internet data providers such as Alexa to measure how active they are and how popular they might be. We match the domain name data with Alexa data to get further information about the popularity of different domains. These exercises are discussed in detail later in the paper.

Having collected this data, we used simple machine learning techniques to develop predictions about the growth potential of each domain name, using which we assigned each domain name a "score". Here we build heavily on methods developed in Guzman and Stern, 2015. Such methods have been shown to work well in other contexts, and we are hopeful of providing one of the first applications of machine learning techniques for website classification.[1] What is novel in our context is that we match domain name data to data from Mergent Intellect, a market intelligence company that tracks companies and their levels of sales and employees and also from Crunchbase, a popular dataset of self-identified "startup" companies. We used Mergent and Crunchbase data to machine-learn what aspects of a domain name record predict growth on a subsample of the data, and then used the results of our learning for "out-of-sample" prediction.

The final output of this process is a filtered list of domains and firms classified according to growth potential score. This growth score provides a way for policy-makers to assess how active entrepreneurship is at any given point in time, at an arbitrary level of geographic specificity. We believe this to be an extremely powerful technique because it promises to encompass entrepreneurship at a global level using a centralized and uniform methodology.

Figure 1 provides a flow-chart of our overall methodology and the following sections illustrate each part in greater detail.

---

1 See http://burning-glass.com/ and http://growthintel.com/ for related applications of this methodology.

2.2 Geocoding

Our first step was to acquire data on WHOIS registration records from WHOISXMLAPI for the UK. Our data consisted of 6,421,689 whois records from the UK, each including fields such as domainName, registrarName, contactEmail, whoisServer, nameServers, createdDate, registrant address, and administrativecontact address. Having obtained these records, our first task was to restrict our analysis the Cambridge and Oxford and geocode the addresses for each of the domains so that we could more precisely identify the geographical location of each of the domain name registrants. We did this as follows.

First, it is useful to note that the UK whois API records contain the registrant addresses, which includes four street address fields (registrant_street1, registrant_street2, registrant_street3, registrant_street4), a city field (registrant_city), and a postcode field (registrant_postCode).
Similarly, the records also contain administrative contact addresses with four street address fields (administrativeContact_street1-4), a postcode field and a city field. None of the UK records seem to have a billing contact address, technical contact address, or zone contact address; only domain registration records from the US have these fields.

When geocoding for the records from Cambridge and Oxford, we first looked at all records where the registrant city is Cambridge or Oxford. To obtain these records, we included all the records where the registrant_city field is equivalent to "cambridge" or "oxford" when all the letters are made lower case. For Cambridge, there were 34839 records while for Oxford, we found 24500 records. Next, we geocoded the addresses for these domains through Arcmap and were able to successfully code about 50% of the addresses. For the rest, we re-geocoded those addresses with the OpenCage Geocoder (see opencagedata.com), who generously offered to let us geocode unlimited addresses with their API. For about 2% of the total data that still did not geocode successfully with the OpenCage API, we geocoded them again with QGIS, which uses Google Maps to match the addresses. There were still a couple dozen addresses that were not matched: these addresses are likely to be written inaccurately, and this is probably the most comprehensive geocoding result we will be able to get. See Figure 2 for a map that uses these geocoded addresses to plot the location of all domain names registered in Oxford and Cambridge.

2.3 Creating other Indicators from WHOIS data

Having collected the WHOIS data, our next step was to create a few other indicators from these data that would allow us to predict the growth score of a domain name. We began by examining the domain name itself and asking whether it contained keywords that are well-known to be present in the names of growth oriented firms. We utilized Crunchbase surveys and also previous research (Guzman and Stern 2015) to identify a list of keywords that are disproportionately present in the names and therefore the domain names of growth oriented firms. A list of these keywords is provided in Table 7. For example, if the domain name contained the words “technology”, or “mobile” or “systems” – then we estimated that it was likely to correlate with a growth-oriented outcome. For each domain name we coded whether one of these keywords was present in the domain name. In addition, to account for the fact that the webpage of many tech blogs and personal websites of professionals also include a high frequency of these keywords, we also created a list of “bad keywords” including words like “I” and “blog”. See Table 7 for a full list. Using the webpage data and the wayback machine data, and the list of “bad” keywords in Table 7, we created the variables scraped_badkeywords and wayback_scrapedbadkeywords.

We also calculated the “length” of the domain name as an independent predictor of growth, because it has been posited that companies with longer names are more likely to be growth oriented. Further,we also looked for eponomy. Previous research has shown that when firms are named after their founders (Chatterjee et al 2015) they are more likely to be associated with a growth oriented outcome. We used the registrant name and compared it to the domain name through a fuzzy matching algorithm and tried to estimate whether a domain name was eponymous. Finally, we obtained the geographic locations of both the University of Oxford and the University of Cambridge, and calculated the distance between the university and the geocoded address for each domain name in order to estimate whether a domain name registrant was physically "close" to a university or not. We hypothesized that domains that were registered in areas closer to the university were more likely to be growth oriented.

2.4 Collecting and Combining WHOIS data with Other Data Sources

One particular unique feature of using domain registrations as a data source is our ability to look at not just the records themselves but also the respective websites they represent. To utilize this innovative source of data, we scraped each individual domain's homepage. That is, we wrote python programs that accessed the homepage of each domain and downloaded the text for each page. Having obtained this data, we then looked at the text of each page to estimate the presence of keywords outlined in Table 7. For each domain we calculated the number of times these keywords were present in the websites (creating the variable scraped_keywords).

For domains that were no longer active, we also scraped the latest archived website of each domain using wayback machine (archive.org). The wayback machine is an internet service that records snapshots of websites and allows for

researchers to get data from a time in the past.
We typically obtained snapshots of the homepage a few months after the registration date to get an accurate sense of the technology sector in which the domain name was operating. Through this procedure we created two new variables: scraped_keywords and wayback_scrapedkeywords.

Finally, we exploited the fact that domain names are often tracked by internet data providers for other purposes to estimate their popularity and the type of their content. In particular we relied on Amazon Alexa, an internet service that routinely tracks "top" domain names and provides estimates of traffic with a relative "rank", as well as an estimate of the likely category that a domain name belongs in, for example "technology", or "media" or "sport". We used the Amazon Alexa Web Information Service (AWIS) for each domain name in our data set to get three additional variables, the "rank", the "usagestats" (i.e. an estimate of how many visitors a domain name receives) and the "category" for a domain name. This was only a preliminary exercise, but there is the possibility that this effort could be expanded to contain significantly greater information about domain names from many other data providers.

Ultimately, these three additional data sources – the content of the page, the archived content of the page from the wayback machine and data from Alexa –

are important to our model, and also illustrate the significant possibilities of this method to be extended and improved. Combined the seven new variables from these three data sources are important features in our machine learning model.

2.5 Validating our Data

After counting all the listed domain registrations in the UK data, we got approximately 6.4 million records. About 1.45 million out of the 6.4 million records are from London (about 20%). This is reasonable because although London has only about one eighth of the total UK population, forms among the largest concentration of higher education institutes in Europe and it contributes about 20% of UK's GDP. Cambridge and Oxford had around .04 and .03 million registrations and account for .6% and .4% of UK's total number of registrations respectively. This is expected due to their much smaller population and economy compared to London. However, since they are both university towns with highly educated populations and centers for technological activities, it is reasonable that the number of domain registrations is still really high compared to the overall population of UK.

Comparing Our Data with Crunchbase

In order to validate the comprehensiveness of our data, we compared our list of domain registrations with the list of homepage domains from Crunchbase, a database of the startup ecosystem consisting of investors, incubators and start-ups.

Of all the organizations under Crunchbase, there were 14310 that are listed under Great Britain. However, only 6356 of those 14310 homepage domains are found within our database of UK domain registrations, which suggests that our list of data left out over half of the domains that are listed as UK companies on Crunchbase.

Further search showed that 3822 of the domains on Crunchbase were .UK domains, which were specifically not included in our dataset, making it completely reasonable that they would not be matched. Taking into account of this, there were still 4132 domains (almost 1/3) listed as UK startups on Crunchbase, but is not included in our dataset.

In order to get a better grasp of the problem, we looked up a random sample of the not-matched domains on whoisxmlapi.com. It seems that one major problem that led to the domains not being matched is that a number of the domain registrations are private, registered through registrars such as domainsbyproxy, perfect privacy Inc., myprivateregistration, contact privacy Inc., and oneandone private registration. Domains registered under these domain name registrars end up having a generic address, which will be completely different from the registrant's actual address. We estimate that these private registrations account for around 60% to 80% of the un-matched domains under Crunchbase.

Another problem is that some of the domains may have out of date information in their Whois record. For example, bitstamp.net listsSlovenia as their country even though it moved its company registration to the UK in 2013. Others such as mangob2b.com put Hong Kong as its address. It does have an office in Hong Kong, though its main office is still in the UK. We estimate that about 20%-30% of the unmatched addresses have this problem. This issue offers both a challenge and an opportunity. On one hand, if digital firms are typically founded in one region but migrate to another, then our index will fail to pick up firms that are relevant for analysis. On the other hand, this database offers a wonderful opportunity to study "entrepreneurial mobility", the likelihood of firms from one region migrating to another -- another important phenomena for which we currently lack systematic data.

These validation errors pose significant problems to our method. In particular, issues around private domain registrations and irrelevant addresses are challenging and hard to address. However, given their small percentage, we ignore this issue for now and focus on the majority of the domains that are included in our database. Also because we do not include .uk and .co.uk addresses in our current methodology, we expect that once these domains are added, the match rate should increase substantially. However, users of this method should be cautious that this challenge could be more severe in certain settings as compared to ours.

2.6 Training Data

Having validated our data, we turned to collecting "training" data to bootstrap our machine learning model. In particular, we aimed to collect data on firms in the Cambridge and Oxford area that we knew achieved a growth outcome. We then matched these firms to the domain name in our records and the list of covariates that we generated (such as the count of keywords, eponymy, Alexa rank etc). We then used a subsample of our complete data (about 30%) to learn which factors predicted a growth outcome as measured by our training data, and we then used this "learning" to predict growth outcomes for our hold-out sample (about 70%) of our data. The training data were also useful because they could be used to validate and test the results of the holdout sample.

The first data source we used for the training data was Mergent Intellect, a flexible web-based application that features a large collection of worldwide business information, including information for 10942 companies based in

Cambridge and 7141 companies based in Oxford. In order to measure whether a company had achieved a growth outcome, we focused on firms that had either greater than five employees or at least 0.5 million dollars in revenue. Conveniently for us, the Mergent Intellect data also included information on the domain name of the company, so we used this field to match the Mergent Intellect data to our dataset. A domain was marked as having achieved a growth outcome (an indicator variable) if it features in this filtered list of Mergent Companies (i.e. if there is a match).

Despite Mergent being a more comprehensive list of companies, we found that there were still a number of domains of promising tech or finance companies which scored high according to our model but were not included in Mergent Intellect. Some examples include Agnito (agnitouk.net), a technology company that help design, implement and maintain computer and audio visual systems and Alchemyst (alchemyst.biz), an E-commerce company. Other high scoring domains included personal websites of professionals, blogs of finance or tech related issues, or social platforms for professionals such as STM (stmassoc.org), a leading global trade association for academic and professional publishers and MedComms Networking (medcommsnetworking.com), which facilitates networking and dialogue amongst individuals working in and around the pharmaceutical industry.

In order to mitigate these concerns of “false negatives” we also matched our domain name data to Crunchbase data for Cambridge and Oxford. In all of our analysis we defined a positive growth outcome as either appearing in Mergent and having achieved greater than 5 employees or $0.5 million in revenue or appearing in the Crunchbase data.

3. Machine Learning and Regression Results

3.1 Summary Statistics

Having assembled the data for our method, we started the regression analysis. Table 1 (Oxford) and Table 4 (Cambridge) provide the summary statistics for all the data in our study.

There are three categories of variables – the ones that come from the WHOIS data itself, the ones from web-scraping and the third from Alexa. For each variable in these three categories Tables 1 and 4 provide summary statistics. “Company=0” is the subsample when a domain name is not identified to be a associated with a growth outcome and does not match with a firm as identified by the Mergent data, while Company=1 is the subsample when the domain names do match Mergent companies.

As Table 1 illustrates, the Oxford dataset contains about 24202 domain names. Of these, 19925 are identified as “company=0” while the remaining 4277 are identified as “company=1”. As can be seen, domains that are associated with companies have higher average probability of having a ‘good’ keyword (0.87 as compared to 0.56), and are slightly longer (13.7 words vs. 13.1). They are also much less likely to have “bad” keywords (0.13 vs 0.18) and have significantly higher pageviews per million (0.23 vs 0.15).

Similarly, as Table 4 illustrates, the Cambridge dataset contains about 24379 domain names. Of these, 20472 are identified as “company=0” while the remaining 3907 are identified as “company=1”. Again, domains that are associated with companies have higher average probability of having a ‘good’ keyword (0.2 as compared to 0.1), and are slightly longer (13.8 words vs. 13.1). They are also much less likely to have “bad” keywords (0.20 vs 0.17) and have significantly higher pageviews per million (0.24 vs 0.12).

These summary statistics help understand the data and provide confidence that the covariates we have assembled are indeed predictive of growth outcomes for domain names.

3.2 Regression Analysis

We now proceed with trying to predict the probability of a domain name being associated with a growth outcome based on the different covariates that we have assembled. We use logit models for this part of the methodology.

Table 2 provides the results of the logit model for Oxford where a growth outcome is defined as a company appearing in the Crunchbase data, while in Table 3 the growth outcome is defined as achieving at least 5 employees or $0.5 million in sales as indicated by Mergent Intellect. In the results we will discuss the results from the Mergent Intellect data, even though the explanation would be similar for the Crunchbase data as well.

The results from Table 3 are illustrative. Column 1 includes variables only from the WHOIS data, Column 2 adds data from the scraped webpage data (include current pages and the wayback machine) and Column 3 adds Alexa data as well. It is quite clear for the positive and significant sign on the "keyword in domain" variable, that this measure is a strong predictor of a growth outcome. Similarly, other variables that positively predict growth outcomes include eponymy, being close to a university, keywords in the domain, and pageviews and Alexa category. Some variables do not have significant predictive power when all variables are added in Column 3, including the Alexa rank and the length of the domain.

Table 6 presents similar results for the city of Cambridge, UK. In Cambridge as for Oxford, we find that keywords, pageviews and the category has significant predictive power for the likelihood of a domain achieving a growth outcome. However, there are some important differences. Eponymy does not seem to be as important in Cambridge as compared to Oxford, and the length of the domain does seem to predict growth outcomes in Cambridge, while in Oxford it does not. This exercise illustrates that different places might have slightly different characteristics that make it important to train the model according to local factors.

These differences between Oxford and Cambridge are important to take note of. Despite their similar cultures, history and economy, we find that the performance of our algorithm differs to some extent. This suggests that our methodology might need to be tweaked further when applied to geographies further afield than the UK. Our suggestion would be to pick training locations in the same country or state as the target region, and then extrapolate the algorithm to other regions in the same country or state, rather than applying the model from one country to another. Given the differences we see even among similar cities such as Oxford and Cambridge, we expect such a customization strategy to be more effective in detecting entrepreneurship across regions.

With that caveat however, broadly, both Table 3 and Table 6 illustrates the power of this approach – namely that using a small number of variables collected from the public internet, it is possible to predict growth outcomes of firms.

3.3 Results (Scatterplots and Prediction Errors)

Figure 3 validates our machine learning and regression algorithms by showing that domain names that rank highly in our algorithm in terms of the score, are significantly more likely to be identified as firms with either 5 or more employees or greater than $0.5 million in revenue in the Mergent Intellect data. In this chart, domain names are ranked based on their score in our algorithm in the X-axis, and their score is on the Y-axis. In other words, domain names are sorted in descending order of their score and plotted in this scatter plot. For domain names that are identified to be firms that satisfy the funding or the employment threshold in the Mergent Data are marked in blue, while the others are identified as red.

As is evident from both the Cambridge and the Oxford data, firms that rank higher according to our analysis in terms of growth score, are also significantly more likely to be companies according to the Mergent data as indicated by the concentration of blue points to the left of the chart. These scatterplots therefore validate our methodology because firms that rank higher according to the predicted score are also more likely in the data as described by the visual evidence.

Another way in which our methodology could be validated would be to see “false positive” and “false negative” as defined by certain growth score thresholds, where “false positives” are firms that score highly in our methodology but which are absent in the Mergent data and “false negatives” are firms that are identified to be “companies” in the Mergent data, but which score low according to our methodology.

Table 8 reports the results from such an analysis. Panel A shows the results from the Oxford data and Panel B shows results from Cambridge. When we limit our sample to the top 25% of the ranking according to our growth score, we find that about 26.97% of these firms are in the Mergent data and the top 25% covers about 40% of the total Mergent sample. This indicates that our procedure is able to successfully predict a growth outcome at a rate that is better than random. A purely random process would mean that of the top 25% of the sample would pick out the top 25% of the Mergent sample, but this estimate is significantly larger. As we make our sample more restrictive, our rate of false positives decreases, but the rate of false negatives grows substantially longer. For example, when limit to only the top 1% of our sample, almost 40% of our domains match with firms in the training data, however, of the top sample, we are including only the top 2.5% of firms. The estimates for the Cambridge data are also quite similar.

Overall, this analysis shows that our rate of false positives and false negatives is still quite significant, which shows that our procedure needs to be revised and improved in order for it to be effective in the field. Having stated this limitation, it is important to remember that we are using only a few variables for prediction, and going forward it would be possible to significantly expand the set of variables in order to improve the estimation. For example, commercial machines learning algorithms commonly employ more than a 100,000 variables for training, while our models use only a handful. If our analysis was expanded in this way, we are quite confident that our rate of Type I and Type II errors would be significantly reduced. Another limitation of our study is that our method is only as good as our training and validation data, and we believe that the Mergent Intellect data is overly restrictive and does not include many companies that one might like to include in the final analysis. This could also explain the high rate of false positives and negatives in our results in Table 8. Future work should pay attention to adding more variables and also utilizing accurate and comprehensive training data to better validate the machine learning exercise.

4. Conclusions

Overall, we provide a novel method to use domain name data to understand entrepreneurial activity. We showed how one might collate domain name registration data, geocoded data about firm location and a host of other variables (including scraped webpage data, Alexa information and information from the Wayback machine) to help predict entrepreneurial activity. We showed how one might use simple machine learning models to help predict startup activity in any given region. We were able to validate our methods using data from Mergent Intellect and show that we are able to successfully predict whether a domain name will achieve a growth outcome, i.e. achieve 5 or more employees, or greater than $0.5 million in revenue.

Our method offers numerous benefits. The most important advantage for using domain registrations as as a data source is that it provides an easily accessible and comprehensive set of data. Most previous research on this entrepreneurial quality rely on data that are limited in scope and not widely accessible such as using business registration records. Validating domain registration records as a source of data will allow anyone interested in researching entrepreneurship, startups, or a similar field to easily gain the necessary data from any country or city in the world at little to no effort. Further, domain name data methods are centralized and global – which means similar methods can be applied all over the world and over time, allowing for simple comparisons across regions and over time. Finally, our methods are relatively inexpensive and do not rely on expensive surveys and hand-collection of data, which make them especially appealing.

Having specified the benefits, it is important to keep some limitations in mind. Most domain names represent non-existent websites or personal websites, and the largest obstacle to getting accurate results from our data set is filtering out the less useful records from the useful records. Another problem is that not all registrants put accurate information in their domain registration and also, an increasing amount of private domain registrations, which allow registrants to hide their contact information and address also makes it hard to get comprehensive results from just the domain registration records. Also, some registrants prefer to keep their information private which also makes this method a challenge. Our methods also currently suffer from a high rate of false positives and negatives, which could be improved on using additional data and more sophisticated machine learning models.

In conclusion, we are very excited about opening up a new avenue for the measurement of entrepreneurship. We are hopeful that the approach that we present will offer policy-makers are more timely, accurate and comprehensive source of data to measure regional entrepreneurial performance. Further, our index relies on fundamentally open data and tools, which would make it possible for independent agencies and sources to evaluate regional new business activity, another important goal. More generally, we're looking forward to the possibility that domain name registration could become one important indicator through which global entrepreneurship is measured.

5. Tables and Figures

Figure 1. Flowchart of the Methodology and Data Sources

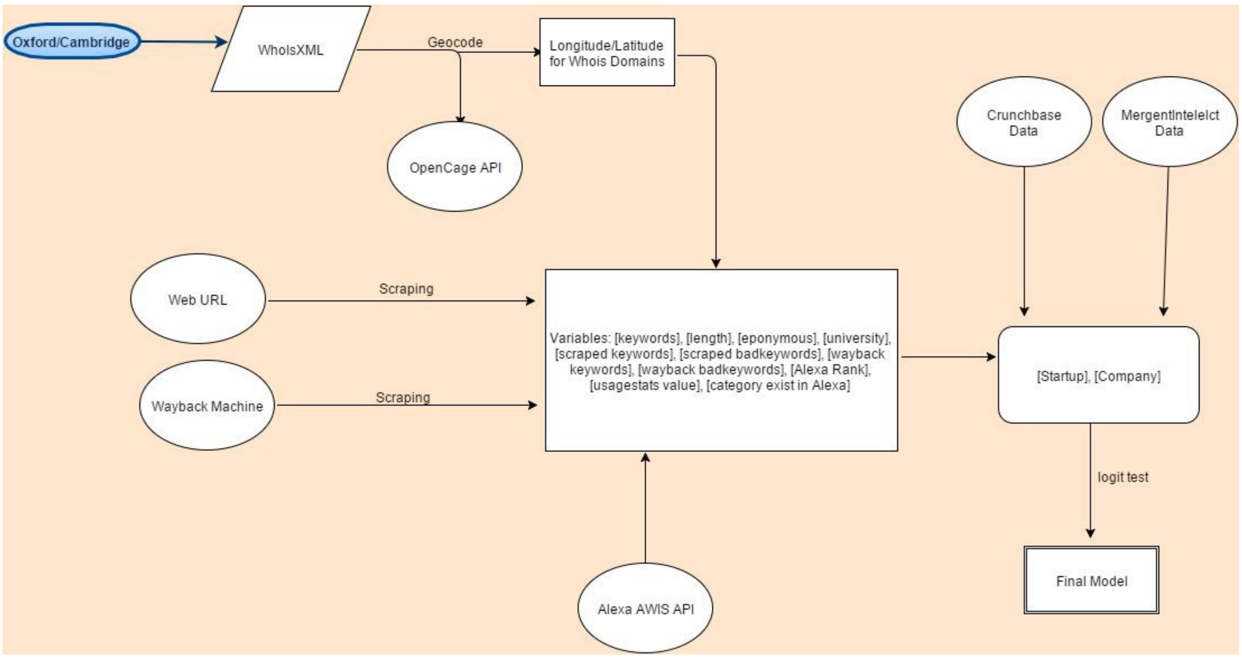
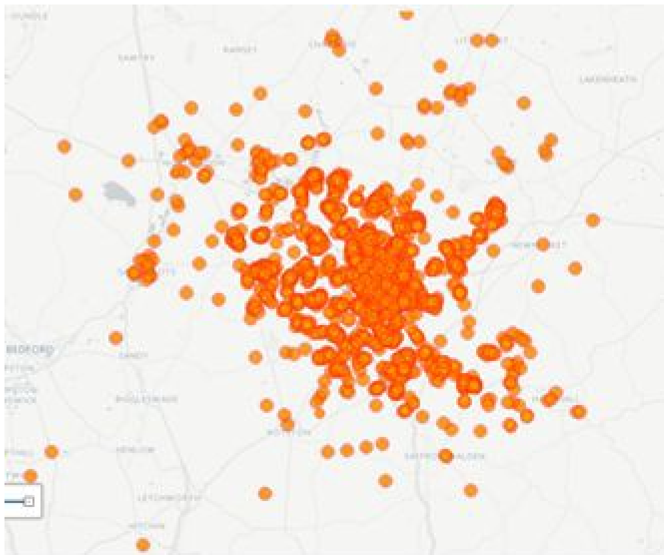


Figure 2.

Panel A : Cambridge



Panel B: Oxford

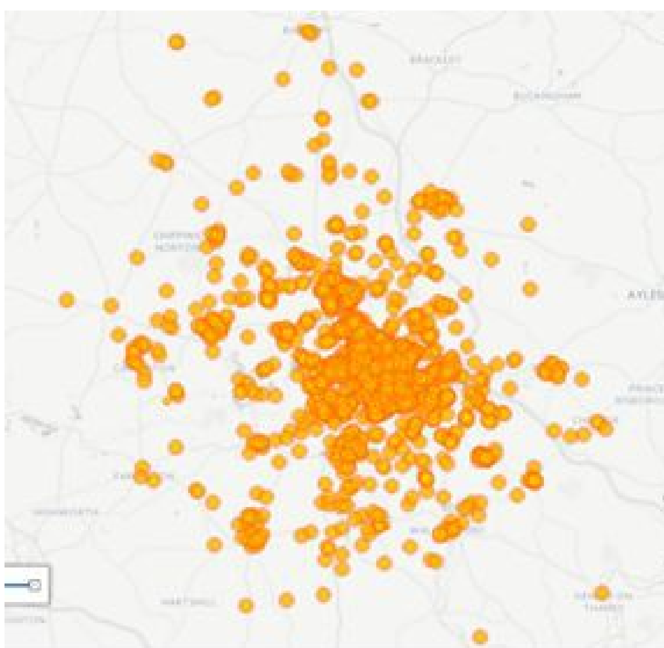
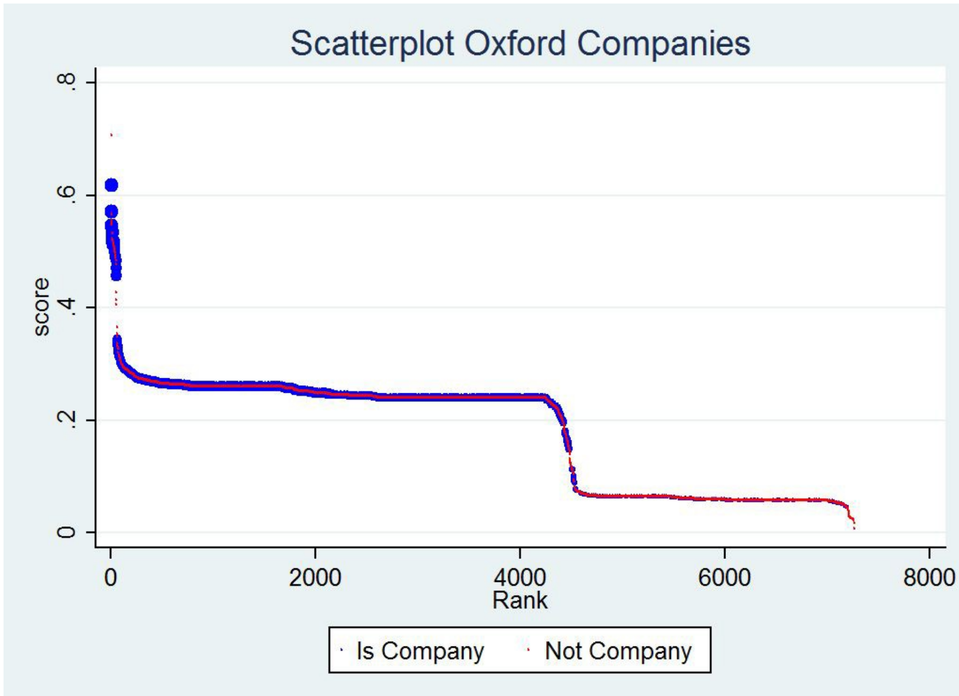


Figure 3.


Scatterplot Oxford Companies
score
0
.2
.4
.6
.8
0
2000
4000
6000
8000
Rank
Is Company
Not Company

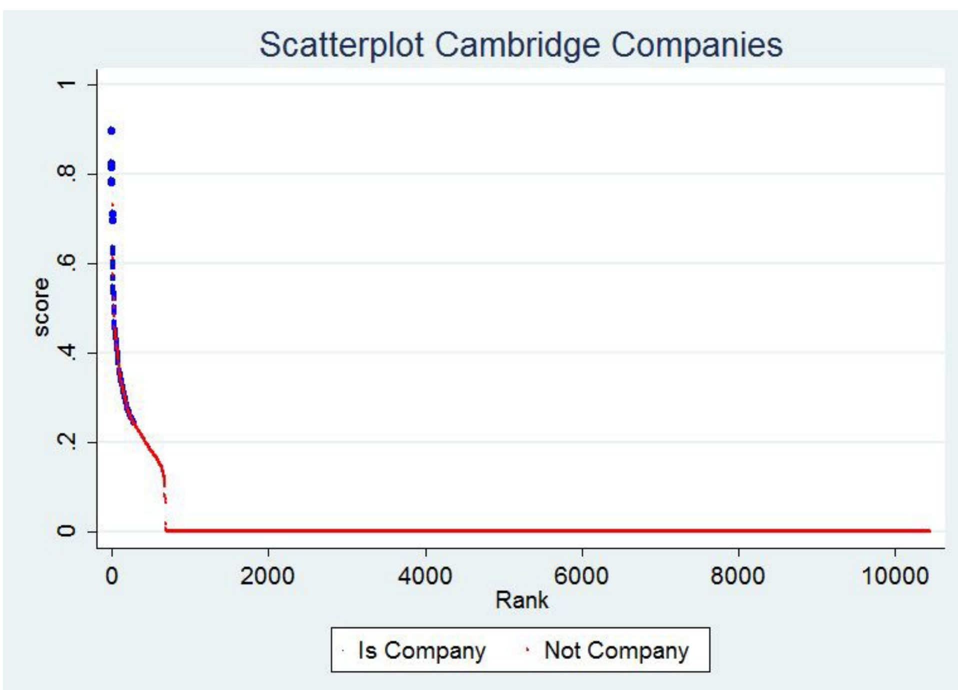

Scatterplot Cambridge Companies
score
0
.2
.4
.6
.8
1
0
2000
4000
6000
8000
10000
Rank
Is Company
Not Company

TABLE 1. Summary Statistics for Oxford Domains

	All Domains			*Company=0*			*Company=1*		
	N	**Mean**	**Std. Dev.**	**N**	**Mean**	**Std. Dev.**	**N**	**Mean**	**Std. Dev.**
Domain Registration Information									
keyword in Domain	24202	0.6181721	0.4858447	19925	0.5630615	0.4960198	4277	0.8749123	0.330857
Eponymous	24202	0.0093381	0.0961834	19925	0.0106399	0.1026022	4277	0.0032733	0.0571259
Length of Domain	24202	13.11218	5.152196	19925	12.98068	5.051768	4277	13.72481	5.556098
Near University	24202	0.8041484	0.396863	19925	0.7850439	0.4108022	4277	0.8931494	0.3089593
Machine Learning									
Active	24202	0.2885299	0.4530881	19925	0.2872773	0.4525034	4277	0.2943652	0.4558102
Scraped Keywords from Domain	24202	1.119701	2.297528	19925	1.085721	2.247616	4277	1.277999	2.511268
Scraped BadKeywords from Domain	24202	0.1784563	0.8759288	19925	0.1878043	0.9168731	4277	0.1349076	0.6504583
Scraped Keywords from Archived	24202	0.9061235	4.76479	19925	0.8429109	4.939466	4277	1.200608	3.834689
Scraped BadKeywords from Archived	24202	0.1236261	0.7785212	19925	0.1280803	0.810155	4277	0.1028758	0.6095465
Alexa Web Information									
AlexaRank<500,000	24202	0.0386745	0.192822	19925	0.037591	0.1902096	4277	0.0437222	0.2045003
Page View Peruser	24202	0.0024151	0.1234676	19925	0.0024793	0.1350723	4277	0.002116	0.0356083
Page View PerMillion	24202	0.1719564	1.061751	19925	0.1580527	1.073175	4277	0.236729	1.004413
Category According to Alexa	24202	0.0135526	0.1156265	19925	0.0099875	0.0994394	4277	0.0301613	0.1710511

TABLE 2. Logit Regression on Crunchbase Data Oxford

		1	2	3
	Keyword in Domain	0.764	0.503	0.423
		[0.393]	[0.401]	[0.403]
	Eponymous	0	0	0
		(empty)	(empty)	(empty)
	Length of Domain	-0.119+	-0.107*	-0.088+
		[0.0401]	[0.0399]	[0.0395]
	Near University	0.337	0.273	0.368
		[0.485]	[0.496]	[0.509]
Machine Learning				
	Active		-1.383	-1.251
			[0.740]	[0.744]
	Scraped Keywords from Domain		0.238**	0.236**
			[0.0347]	[0.0366]
	Scraped BadKeywords from Domain		-0.460	-0.388+
			[0.237]	[0.0233]
	Scraped Keywords from Archived		0.0097	0.0108
			[0.00947]	[0.009]
	Scraped BadKeywords from Archived		0.2498**	0.245*
			[0.087]	[0.0945]
Alexa Web Information				
	AlexaRank<500,000			1.62**
				[0.476]
	Page View Peruser			-0.0726
				[0.510]
	pave View PerMillion			0.0827*
				[0.0292]
	Category According to Alexa			1.757**
				[0.484]
	Observations	23976	23976	23976
	Pseudo R-squared	0.0243	0.1295	0.1771
	Log-Likelihood	-250.71	-223.67	-211.45

Exponentiated coefficients; Standard errors in brackets + p<.05 * p<.01 ** p<.001

TABLE 3. Logit Regression on Mergent Intellect Data Oxford

		1	2	3
	Keyword in Domain	1.730**	1.718**	1.715**
		[0.049]	[0.049]	[0.049]
	Eponymous	0.962**	0.952**	0.963**
		[0.282]	[0.282]	[0.283]
	Length of Domain	0 .0036	0.004	0.005
		[0.0033]	[0.0033]	[0.0033]
	Near University	0.954**	0.957**	0.964**
		[0.053]	[0.053]	[0.053]
Machine Learning				
	Active		0.0952+	0.109*
			[0.039]	[0.039]
	Scraped Keywords from Domain		0.030**	0.255*
			[0.008]	[0.008]
	Scraped BadKeywords from Domain		-0.13**	0.123**
			[0.0355]	[0.0357]
	Scraped Keywords from Archived		0.0157*	0.0129*
			[0.005]	[0.005]
	Scraped BadKeywords from Archived		-0.010	-0.0279
			[0.036]	[0.037]
Alexa Web Information				
	AlexaRank<500,000			0.146
				[0.088]
	Page View Peruser			-0.210
				[0.298]
	Pave View PerMillion			0.0409*
				[0.014]
	Category According to Alexa			1.048**
				[0.126]
	Observations	24202	24202	24202
	Pseudo R-squared	0.0910	0.0931	0.0968
	Log-Likelihood	-10260	-10236	-10194

Exponentiated coefficients; Standard errors in brackets + p<.05 * p<.01 ** p<.001

TABLE 4. Summary Statistics for Cambridge Domains

	All Domains			*Company=0*			*Company=1*		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Domain Registration Information									
keyword in Domain	24379	0.123672	0.3292137	20472	0.1075127	0.3097715	3907	0.208344	0.4061761
Eponymous	24379	0.0081628	0.0899804	20472	0.008646	0.0925831	3907	0.0056309	0.0748375
Length of Domain	24379	13.26314	5.213054	20472	13.14767	5.112513	3907	13.86819	5.67342
Near University	24379	0.7794003	0.4146595	20472	0.7815553	0.4132009	3907	0.7681085	0.4220941
Machine Learning									
Active	24379	0.700644	0.4579854	20472	0.7085287	0.4544512	3907	0.6593294	0.4739954
Scraped Keywords from Domain	24379	0.9593913	2.32653	20472	0.8565846	2.15431	3907	1.49808	3.018747
Scraped BadKeywords from Domain	24379	0.1769966	0.8414185	20472	0.1716979	0.7902328	3907	0.2047607	1.070025
Scraped Keywords from Archived	24379	0.639977	1.583628	20472	0.5739547	1.414801	3907	0.9859227	2.240313
Scraped BadKeywords from Archived	24379	0.0986095	0.5868966	20472	0.098476	0.5804642	3907	0.0993089	0.6195873
Alexa Web Information									
AlexaRank<500,000	24379	0.0033225	0.0575467	20472	0.0024424	0.0493611	3907	0.0079345	0.0887329
Page View Peruser	24379	0.0037865	0.2432048	20472	0.0022015	0.086276	3907	0.0120911	0.5745112
Page View PerMillion	24379	0.1435678	0.7131941	20472	0.123286	0.6609155	3907	0.2498413	0.9337088
Category According to Alexa	24379	0.0115673	0.1069299	20472	0.0082063	0.0902185	3907	0.0291784	0.1683279

TABLE 5. Logit Regression on Crunchbase Data Cambridge

		1	2	3
	keyword in Domain	0.187	-0.0138	-0.397
		[0.381]	[0.399]	[0.460]
	Eponymous	0	0	0
		(empty)	(empty)	(empty)
	Length of Domain	-0.119**	-0.125**	-0.097*
		[0.0300]	[0.0319]	[0.032]
	Near University	-0.203	-0.0563	-0.228
		[0.291]	[0.303]	[0.316]
Machine Learning				
	Active		1.909**	1.43*
			[0.564]	[0.558]
	Scraped Keywords from Domain		-0.0663	-0.061
			[0.0527]	[0.0587]
	Scraped BadKeywords from Domain		-0.158	-0.1677
			[0.216]	[0.232]
	Scraped Keywords from Archived		0.368**	0.3246**
			[0.509]	[0.0564]
	Scraped BadKeywords from Archived		0.0493	0.0169
			[0.162]	[0.1755]
Alexa Web Information				
	AlexaRank<500,000			1.155+
				[0.559]
	Page View Peruser			0.133
				[0.1936]
	Page View Value			0.3167**
				[0.0456]
	Category According to Alexa			1.439**
				[0.3833]
	Observations	24180	24180	24180
	Pseudo R-squared	0.0216	0.1947	0.2813
	Log-Likelihood	-422.51	-347.73	-310.36

Exponentiated coefficients; Standard errors in brackets + p<.05 * p<.01 ** p<.001

TABLE 6. Logit Regression on Mergent Intellect Data Cambridge

		1	2	3
	keyword in Domain	0.744**	0.679**	0.684**
		[0.0459]	[0.0465]	[0.0467]
	Eponymous	-0.386	-0.369	-0.404
		[0.228]	[0.231]	[0.233]
	Length of Domain	0.0199**	0.0229**	0.0245**
		[0.2277]	[0.003]	[0.003]
	Near University	-0.038	-0.0193	-0.0278
		[0.0319]	[0.0424]	[0.0426]
Machine Learning				
	Active		-0.398**	-0.4366**
			[0.0398]	[0.0401]
	Scraped Keywords from Domain		0.0848**	0.0856**
			[0.008]	[0.008]
	Scraped BadKeywords from Domain		-0.0009	-0.00162
			[0.0241]	0.0241]
	Scraped Keywords from Archived		0.085**	0.0655**
			[0.0112]	[0.0115]
	Scraped BadKeywords from Archived		-0.117*	-0.140**
			[0.040]	[0.04098]
Alexa Web Information				
	AlexaRank<500,000			0.218
				[0.267]
	Page View Peruser			0.0061**
				[0.0604]
	Page View Value			0.137**
				[0.0217]
	Category According to Alexa			1.026**
				[0.135]
	Observations	24379	24379	24379
	Pseudo R-squared	0.0145	0.0308	0.0365
	Log-Likelihood	-10573	-10398	-10337

Exponentiated coefficients; Standard errors in brackets + p<.05 * p<.01 ** p<.001

TABLE 7. List of Keywords and Bad Keywords Used for Scraping

Keywords				Bad Keywords	
1	Technology	14	Ventures	1	blog
2	Technologies	15	Solutions	2	review
3	Company	16	Software	3	my
4	Fund	17	Customer	4	me
5	Network	18	Product	5	I
6	Market	19	Careers	6	personal
7	Business	20	Executive		
8	Systems	21	Pricing		
9	Mobile	22	Portfolio		
10	Management	23	Start-up		
11	Partners	24	Client		
12	Digital	25	Games		
13	Services	26	Project		

TABLE 8: Prediction Errors in Oxford and Cambridge

Panel A: Oxford – 17.02% of all domains are identified as companies

Rule for Cutoff By Predicted Score	How many percent are Mergent firms?	What Percent of Mergent Firms are
Top 25%	26.97%	39.56%
Top 10%	26.34%	15.45%
Top 5%	31.96%	9.38%
Top 1%	38.89%	2.26%

Panel B: Cambridge – 15.86% of all domains are identified as companies

Rule for Cutoff By Predicted Score	How many percent are Mergent firms?	What Percent of Mergent Firms are
Top 25%	24.69%	38.36%
Top 10%	28.61%	18.03%
Top 5%	33.46%	10.55%
Top 1%	55.77%	3.50%

6. Bibliography

Guzman, Jorge, and Scott Stern. "Where is Silicon Valley?." *Science* 347, no. 6222 (2015): 606-609.

Haley, Christopher “Launching the European Digital City Index” http://www.nesta.org.uk/blog/launching-european-digital-city-index (2015)

Klapper, Leora, Raphael Amit, and Mauro F. Guillén. "Entrepreneurship and firm formation across countries." In *International differences in entrepreneurship*, pp. 129-158. University of Chicago Press, 2010.

Mandel, Michael. “Using online 'help wanted' ads to track emerging occupations “ http://www.nesta.org.uk/blog/using-online-help-wanted-ads-track-emerging-occupations (2014)

Reynolds, Paul, Michael Hay, William D. Bygrave, S. Michael Camp, and Erkko Autio. "Global entrepreneurship monitor." *Executive Report* (2000).