# Fundamentals of Digital Archeology Final Project Proposal - Analysis of Startup Metrics to Gauge Affect on Success and Failure

L. Mills[1], H. Choi[2], and D. Lowe[3]

***Abstract*— In the following document, an analysis considering factors affecting success and failure of startups is proposed. An observation of startup industry of involvement, date of founding, initial funding, and country of origin will be performed. Herein the proposed comparison of these metrics is discussed.**

## I. BACKGROUND

Startups have been exploding in growth over the past couple decades. They operate in an expansive and diverse range of industries and are formed and operated in a number of locations. Furthermore, many seek funding in their initial operation, further complicating the possible dynamic for fledgling startups.

According to the Bureau of Labor Statistics, 220,000 small companies were formed in the first quarter 2014 while 189,000 were disbanded during that quarter. The rate of startup creation has almost consistently outpaced the rate of startup death over the past two decades, with the exception of the dot-com bubble and Great Recession of 2008. These astounding rates of tech company creation and death leave one in awe, and might lead one to wonder, what factors contribute to the success and failure of a startup? In the proposed project, we look to answer this very question.

## II. HYPOTHESIS

It was hypothesized that initial funding would increase success in startups. Additionally, it is surmised that certain locations will see greater rates of success, while other locations see greater rates of failure. Furthermore, it is predicted that certain industries will result in greater rates of funding - likely the more established industries such as medical, finance, and consumer electronics.

## III. METHODS

### A. Comparison and Metrics

In the study, an analysis of a number of metrics and their affect on the success and/or failure of startup was performed. Though these factors were not exhaustive, they provide a solid baseline off of which various classes of startups might be compared. With this baseline, further study on these or other metrics were performed to establish a more comprehensive understanding on how such subjective metrics affect the success and/or failure of a startup.

The factors which were used for this analysis included the following:

- Country of origin - the country of origin of a sample of startups was analyzed to observe whether or not startups established in certain countries were more likely to succeed or more likely to fail. This could be further expounded upon by studying which regions in particularly selected countries offer greater rates of success, and which offer greater rates of failure. This methodology could be extended to much finer granularity, all the way to studying how rates of success and failure of startups that originate in different districts of particular cities compare.
- Date Founded - the date of foundation for a sample of startups was analyzed. With this metric, the goal was to gain insight into how startups fare across the span of time and in different climates of innovation, investor interest and confidence, consumer confidence, and even economic climate.
- Category of organization - the category of a sample of startups (i.e. organizations) was observed in order to determine whether companies in certain categories are more or less likely to succeed than companies in other categories. Of particular interest was how companies in categories with more established markets (such as medical) would compare to companies in categories of emerging markets (such as autonomous vehicles).

In order to determine how the metrics of interest affect rates of success for startups, the observation of the average funding of groupings of startups that fall in the same class when considered from the perspective of the above metrics will be performed. For example, when analyzing how country of origin affects startups success, the average most recent report of funding for startups in each given country was be computed, then these results were compared across countries. This methodology of comparison was propagated across all the above proposed metrics, with a comparison between classes resulting for each metric. In this manner, the average success of companies in each class for the given metrics was established by observing the resulting average funding for said class, which was then be compared to the remaining classes. It should be noted that companies that have closed will be omitted from this sample, as a company has not succeeded by this definition if it has closed. Furthermore,

*This work was not supported by any organization

[1]L. Mills is a undergraduate of computer science, University of Tennessee, Knoxville Knoxville, TN 37916, USA lmills9@vols.utk.edu

[2]H. Choi is a undergraduate of computer science, University of Tennessee, Knoxville Knoxville, TN 37916, USA hchoi6@vols.utk.edu

[3]D. Lowe is a undergraduate of computer science, University of Tennessee, Knoxville Knoxville, TN 37916, USA dlowe7@vols.utk.edu

companies that have had an initial public offering (IPO) were excluded due to the fact that these companies could no longer be classified as "startups."

In order to determine how the above metrics affect rates of success for startups, the average rates of closure for groupings with respect to the metrics of interest was analyzed. For example, when analyzing how area of industry affects startup failure, the average rate of closure of startups for each given industry was computed. The results of these averages was then be compared across industries. In this manner, the affect of all metrics of interest was compared across classes. In this manner, the average failure of companies in each class was established for the given metrics. As with rates of success, the resulting average rate of closure of companies in each given class was noted and compared to the all other classes for the given metric.

Following the computation of the averages for classes in each given metric, the averages were plotted in such a way to demonstrate the disparity between classes for each metric. Two graph resulted for each of the metrics of interest - one for analyzing the metric's average affect on tech company success, and one for analyzing the metric's average affect on tech company failure.

### *B. Data*

As mentioned, the data for this study was obtained from Crunchbase. Crunchbase is a privately-owned business information platform. As per Crunchbase, "Crunchbase is the destination for discovering industry trends, investments, and news about hundreds of thousands of public and private companies globally." The initial data received from Crunchbase included records of funded companies, funding rounds, investors, acquisitions, and IPOs (initial public offerings). From this, the data was culled down to include only that data of interest, namely the funded companies. This dataset contained highly granular data on individual funded companies, including company name, country, state (if applicable), city, status of the business (open, closed, IPO) category of industry, a list of business sectors participated in, total funding in United States dollars (adjusted for inflation as applicable), date founded, date first funded, date most recently funded, date closed (if applicable), email address, phone number, Crunchbase URL, Twitter URL, Facebook URL, and more. As a finite set of metrics was earlier selected on which to focus for this study, this data was further culled to include company name, country, state (if applicable), city, status of the business, category of industry, total funding in United States Dollars, date founded, and date closed (if applicable).

The data originally included information on companies which had IPO'd, companies of which we were not interested. As such, we eliminated all records of companies that had IPO'd. Furthermore, the data did include empty fields were data was not available or not applicable. As such, incomplete data which rendered a companies data irrelevant was eliminated completely from the data. Particularly, all companies with no record of total funding were eliminated from the study, as this was a key metric used in all parts of our study. Furthermore, as the study was only interested in companies founded since 1990, company records of all companies founded before the first of January in 1990 were eliminated. This reduced the dataset down to 68,627 records of individual companies.

A similar methodology had to be applied to the data for the computation of particular metrics which used various fields of the data. For metrics computed involving the category of industry, the records of companies which had missing data in the category of industry field were eliminated. This reduced the dataset used here down to 66,436 entries. Likewise, for metrics computed involving country of origin, the records of companies which had missing data in the country of origination field were eliminated. This reduced the dataset used in computations involving the country of origin down to 67,259 records of companies. Finally, for metrics computed involving date of founding, the records of companies which had missing data in the date of foundation field were eliminated. This reduced the dataset used in computations involving the date of foundation down to 68,627 company records.

### *C. Context*

The data attained in this study was limited to that available via Crunchbase. Crunchbase provides an admirable record of 152,492 companies, covering everything from the company name to number of employees. While this is formidable, the data which Crunchbase provides not all-encompassing of all startups formed worldwide. In the search for resources, however, it was the best in terms of the shear quantity of startups on which it had data. Furthermore, the breadth of information the data covered could not be matched.

Given the nature of Crunchbase's dataset this analysis is limited to reveal information relating those startups included in the dataset. Though not perfect, the analysis surely reveals insight on a subset of the startup economy and culture. In performing this analysis, this fact was closely kept in mind. As such, while the results presented here are taken to certainly have merit and meaningful implications, the caveat of incompleteness of data must be emphasized to the reader.

### *D. Tools of Analysis*

Due to the immense size of the data, several different tools were utilized for this study. Initially, the thought was that Microsoft Excel would be enough to analyze the data, but the amount of data proved overwhelming and caused frequent crashing. It was therefore necessary to use Python for data parsing in order to get smaller pieces of data that could then be written to CSV(Comma Separated Values) files and used for making various styles of graphs. For collaboration, files were stored on GitHub and Python results were stored using Jupyter Notebook, allowing for easy changes amongst team members.

### *E. Methodology*

In order to analyze the data, a number of computations had to be performed. Though a similar methodology was

performed to compute the various metrics, specific steps had to be taken in each case to adjust for various factors. A good example of this was the normalization in computations performed which involved the year and average total funding. Were the normalization not to have been completed, the averages would have been much higher for past years, as those companies founded, say, for example, ten years ago would have had a longer time to acquire funding, and thus the averages would always be higher for those companies than a company that was founded, say, one year ago.

The computations performed included creating arrays for each of the relevant averages and the counts of those various categories for the class in question (for example, for the various countries startups were formed in). For average funding, the data was then iterated over, the total funding for the company in question was added to the array element of the averages array for the pertinent category, and the count of the category the element in question fit into was incremented by one. Following this, the array of averages was then iterated over, with each value being divided by its associated count in the array for counts of the particular categories. In the computations which involved the year of founding, the data was normalized by dividing the resulting value by the number of years since that date of founding (where date of founding was the category of class in question by which the entries of company data were split up by). This methodology was performed for each of the pertinent metrics of interest in order to compute the average funding for various categories in the three classes of interest: date of founding, country of origin, and category of organization.

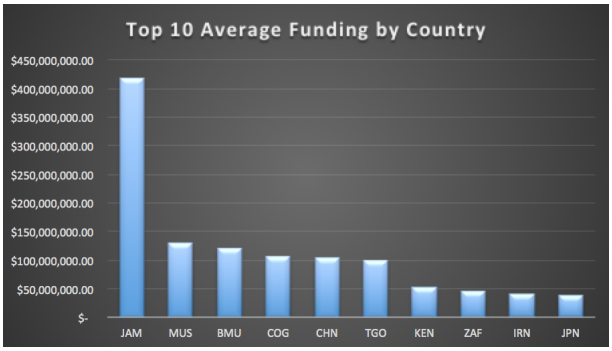
For average rate of closure, the arrays for relevant averages and counts of various categories for the class in question were performed as was the case for the computations of average rates of funding. From this point, the data was again iterated over, with the value one being added to the array element in the averages array for the category that the company in question fit into if said company had closed. Otherwise, nothing was added to the averages array for the category that the company in question fit into. After the data was completely iterated over, the averages were computed by iterating over the averages array (which at this point contained the number of companies closed in that category of class due to our calculations) dividing the value in the array element in question by the associated count in the count array for the respective category which that array element pertained to. Once again, the data was normalized in the computations where the year of foundation was involved. This was again done by dividing those entries of the averages array by the number of years since that date of founding for the particular element.

After these computations were completed, the data was output to a CSV. From this point, the data was handled in Excel. Here, a number of charts and graphs were made in order to interpret the outcomes of the computations. The *RESULTS* sections highlights these findings and interpretations in detail.

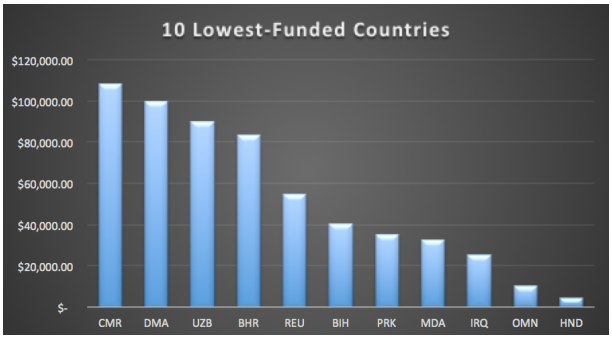
## IV. RESULTS

### A. Funding by Country

When looking at funding by country, it wsa decided that the top 10 highest and lowest funded countries would be observed to see if the original hypothesis that funding was a major factor in startup success would be supported by hard data.



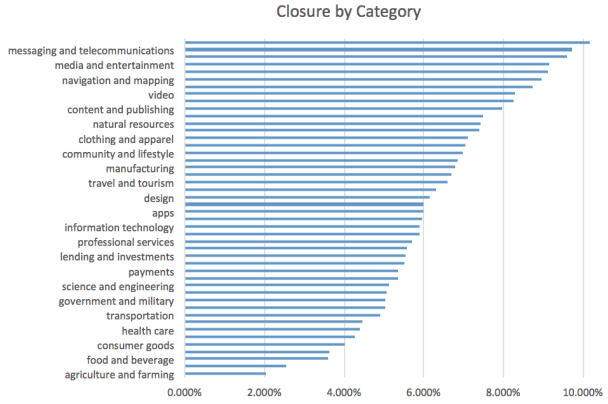

When looking at the highest funded countries, Jamaica stood out at first, and closer inspection would reveal that this was because it only had only three startups since 1990, one of which was funded by a billionaire and would go on to become a major cell phone service provider for the area. Although it was tempting to remove this data entirely for the purpose of looking at the other countries, it was ultimately kept in because despite having so few startups in Jamaica, the closure rate for those was 0%, tying in with the initial hypothesis that funding made a big difference in startups' success rate.




When looking at the lowest funded countries, this matched expectations, comprising countries with low GDP or otherwise stunted technological growth, such as with North Korea.

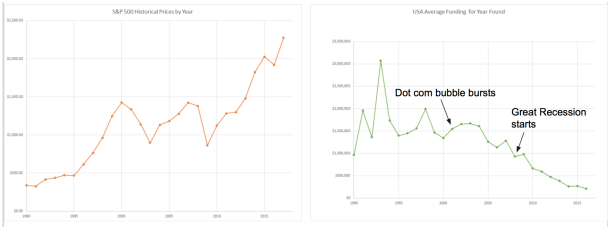
### B. Rates of Closure by Country

When the word ”startup” comes up in today’s vernacular, it typically refers to a tech startup thanks to the current boom going on in Silicon Valley. As a result, the expectation was that tech companies would be doing well when examining rate of closure. However, it turns out that the least likely industries to close were those that were more necessary to survival, such as agriculture, energy, or nuclear energy. Artificial Intelligence startups also appeared to be doing incredibly well compared to other tech fields. The most likely fields to close were various technological fields, which upon further inspection seemed to make the most sense given the rapid pace at which different technologies are being phased out.

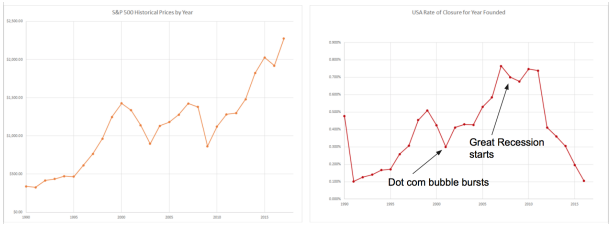



### C. Stock Market Correlation

One interesting correlation noted is that of funding in the US market compared to the rise and fall of the stock market. Notice in the figure below that as the stock market improved, the funding for startups started to trend downward. One possible reason for this is that as earlier startups became more successful over time, this led to a surge in the stock market. The average funding was already declining due to the Dot com burst and the continued decline after the Recessions could be the residual effects of the burst made worse by the dipping economy.




The graph below shows that the startup closure rate in the USA trended upward during the same time period average funding dropped. Specifically closures did climb higher after the Dot com burst, but after plateauing during the Recession it continued to decrease till 2017. These opposing trends may mean that investors are getting more efficient at avoiding bad investments.




## V. Related Work, Limitations, Issues Encountered

### A. Limitations

Limitations included lack of familiarity with statistics tools such as R to get a different look at various data, and also the time to break down various cases more specifically.

### B. Issues Encountered

The data was large and unwieldy, and even though it was provided via spreadsheet, it was so much data that it wasn’t usable without first breaking it up into smaller pieces.

## VI. Conclusions

Unfortunately there was too much data and not enough time to draw a solid conclusion on what factor, if any, influences the success of startups the most. It can be reasonably stated that funding plays a large role in a startup’s initial success, but there are also many factors that cannot be accounted for, such as the drive and ambition of the founder and its team, or even environmental factors that could somehow lead to a closure. As expected, the US had the most startups by a wide margin, and future studies could be done into simply the US alone, perhaps even splitting it up by region and seeing how certain factors specific to the region influenced growth, such as Chattanooga, TN’s push for city-wide Gigabit internet, or the current tech boom going on in Texas right now as people flee the crowded and expensive west coast.

## Acknowledgment

We would like to take the chance here to express our thanks to Professor Audris Mockus, who encouraged us to explore this data and gave us the opportunity to pursue this project, and Luke Mills, who originated the project concept. Furthermore, gratitude is extended to Crunchbase for granting research access to their exceptional system to make this study possible, and for providing their information in an easily accessible manner.

## References

[1] Bls.gov. (2017). Entrepreneurship and the U.S. Economy. [online] Available at: https://www.bls.gov/bdm/entrepreneurship/entrepreneurship.htm [Accessed 29 Sep. 2017].

[2] Bls.gov. (2017). Chart 5. Quarterly establishment births and deaths, 1993-2015. [online] Available at: https://www.bls.gov/bdm/entrepreneurship/bdm_chart5.htm. [Accessed 29 Sep. 2017]

[3] Crunchbase.com. (2017). Crunchbase. [online] Available at: https://www.crunchbase.com/organization/crunchbase [Accessed 1 Oct. 2017]