DRAFT - PLEASE DO NOT CITE

The Link between R&D, Human Capital and Business Startups[1]

Nathan Goldschlag[2]
Ron Jarmin[2]
Julia Lane[3]
Nikolas Zolas[2]

Abstract

We examine the links between startup performance and new measures of workforce human capital. We apply machine learning techniques to a rich new source of longitudinally-linked data to characterize the research experienced workforce of new businesses. Startups with more research experienced workforce are more likely to survive, become successful and more likely to grow.

[1] Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. This research was supported by the National Center for Science and Engineering Statistics. NSF SciSIP Awards 1064220 and 1262447; NSF Education and Human Resources DGE Awards 1348691, 1547507, 1348701, 1535399, 1535370; NSF NCSES award 1423706; NIHP01AG039347; and the Ewing Marion Kaufman and Alfred P. Sloan Foundations. Data were generously provided by the Committee on Institutional Cooperation and its member institutions. We thank Cameron Conrad, Ahmad Emad, Christina Jones and Evgeny Klochikhin, for research support, Greg Carr, Marietta Harrison, David Mayo, Mark Sweet, Jeff Van Horn, and Stephanie Willis for help with data issues, and Jay Walsh, Roy Weiss, and Carol Whitacre for their continuing support. The research agenda draws on work with many coauthors, but particularly Bruce Weinberg and Jason Owen Smith.
[2] U.S. Census Bureau
[3] New York University

Introduction

The past decade has been characterized by a decline in both the formation and the success rate of new firms (*1*); the reasons for the decline are not fully understood. In this paper, we use new measures of human capital to investigate the contribution of human capital composition to declining dynamism. In doing so, we draw on the literature that suggests that economic growth can be significantly affected by workers specialized in R&D (*2, 3*) and that has indirectly shown links between investments in research and innovation (*4–6*). The work complements an extensive literature that links regional economic development clusters with the presence of active research universities (*6–9*). The findings are consistent with the notion that an important source of knowledge transfer is the flows of research experienced workers from one firm to another (*10, 11*).

We incorporate new individual-level measures of R&D human capital, including research training, of the workforce at both startups and young firms to directly examine the connection between an R&D trained workforce and new business success. As such, the paper also demonstrates a new pilot approach to scaling and augmenting existing data collected at a local or regional level or for a subsample of firms and individuals. Using survey data to document the impact of research funds on local and national economic outcomes would both be prohibitive in terms of cost and uncertain in terms of quality. As such, we utilize machine learning to scale our sample and generate estimates of the workforce with research experience.

The results suggest that a one-unit increase in the number of research experience employees in a startup firm's workforce increases its probability of survival to the next period by 1.7%, and increases the likelihood of becoming a high-growth successful startup by 0.8%. Workers with experience in research increase the likelihood of startup success (defined as having survived for 5-years with 10+ employees) by 2.9%, over and above workers who had been employed by universities, High Tech or R&D performing firms.

These results are consistent with the view that there is a relationship between workforce experience and business startup and survival. Further work using these data will be necessary to examine temporal dynamics. It will be particularly interesting to understand whether changes in the fluidity of this type of workforce, changing patterns of firm-to-firm job flows, or changes in the nature of research funding, can be tied to the decline in business dynamism and changes in the distribution of employment growth rates.

Literature

There is a growing body of evidence about the decline in business dynamics in the past 30 years (*12–14*). At least some of this is due to the decline in responsiveness of both young and mature businesses to shocks, which have resulted in substantial changes in the contribution of reallocation to productivity growth (*15*). The findings are evident in all geographic areas as well as in narrowly defined industries. In addition, the distribution of growth rates has narrowed substantially since the 1990's, suggesting the relative returns to success have declined (*15*). While some research suggests that the decline is due to changes in demographics (*16*), the growth rate of labor supply (*16*) or due to the slowing of progress on the technological frontier (*17–19*), the core reasons are not well understood (*1*).

One possibility is that hitherto unmeasured workforce characteristics are contributing to changes in business dynamism. The decision to start a business, and its subsequent productivity and success is associated with having an entrepreneurial workforce (*7, 20*). It is clear that the past 40 years have also been characterized by declining labor market fluidity (*21*). This finding is important since worker reallocation is one way in which economic growth occurs and there is a great deal of worker reallocation in the economy (*22*, *23*). However, it has been difficult to use standard measures of workforce characteristics to explain the changes in business dynamism. Age, sex and marital status have limited explanatory power (*21*), and the standard measure of human capital – years of education – does not change rapidly. However, Kerr et al. suggest that other measures of human capital, such as experimentation, may be important factors (*13*). Related work also suggests that highly innovative individuals make "exceptional" contributions to economic growth (*24*).

There is a new opportunity to develop proxies for such other measures, particularly with the advent of new longitudinally linked datasets. Human capital in such areas as experimentation may be acquired through both formal training and on the job learning. As such, sensible measures might include experience in R&D performing or High Tech businesses. Linked employer-employee data, such as the LEHD data (*25*), can be used to construct such measures. Such data have been used in the past to generate different measures of individual experience at different types of businesses (*22*). Barth et al., for example, show that there are returns to experience at R&D performing firms (*26*); Abowd et al. also use linked data to compute person specific measures of human capital (*27*).

More direct measures of research human capital are now available, which include specific information on whether workers are trained in scientific research. Arguably, the scientific method is the encapsulation of experimentation and refinement in the face of both success and failure. The new longitudinally linked data on the research trained workforce - the UMETRICS data, (*28*) - have been used in other contexts and do suggest that research trained individuals are more likely to work at firms with characteristics closely linked to productivity (*29*).

Approach, Data and Measurement

Our framework posits that startup outcomes (Y) such as the survival and subsequent success of a startup *f* at time *t* is driven by capital and technology (AK), quantity and quality of labor measures (L) such as human capital, and external factors (X) such as macroeconomic conditions and industry factors. Functionally, we can think of outcomes being written as:

$$Y_{ft} = f\left(AK_{ft}, L_{ft}, X_{ft}\right)$$

For firm *f* at time *t*. We construct measures for each of these components using existing Census microdata on linked employee-employer data, longitudinal firm-level data, as well as existing surveys which indicate whether or not the firm is or was an R&D performing firm. We supplement this data with new data from UMETRICS, which identifies all individuals who were paid on research grants for 14 universities that account for approximately 15% of federally

funded research. Our primary focus is constructing components for the measure L_{ft}, which consists of the attributes of the startup workforce at time *t=0*.

Identifying Startups and Startup Outcomes

We create a Startup Firm History File (2005-2014) based on a panel database of age zero establishment attributes. The primary frame for the data is the Longitudinal Business Database (LBD), supplemented with additional information from the Census Bureau's Business Register upon which the LBD is based. We utilize this file to identify startups by yearly cohort. Once the startups have been identified, we supplement the data with geocodes and EINs taken from the Business Register. These variables are used to subsequently characterize the workforce associated with each startup gathered from LEHD (Longitudinal Employee-Household Dynamics) and W2 records. The full file contains data on employment, payroll, industry, geography, firm-type and birth/date of the firm.

Figure 1 below provides a graphical summary of the number of startups each year, including the share that fail in the subsequent years.

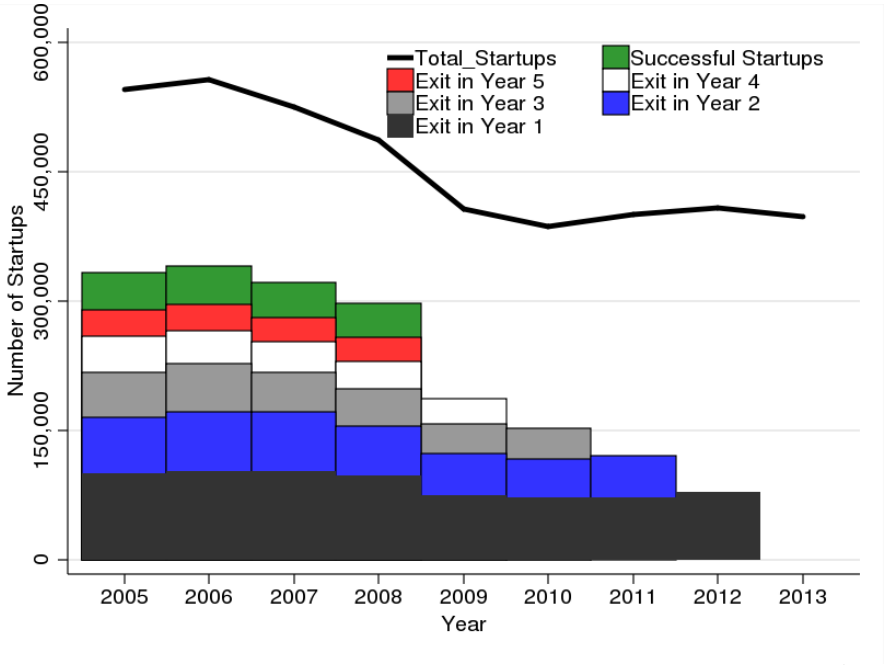

Figure 1: Number of Startups and their Death Rates first 5 years[4]

Figure 1 shows the counts of startups, as well as exits in each subsequent year, for the data sample. It also shows how many "successful" startups there are (defining success as surviving to year 5 and having more than 10 employees). Consistent with earlier findings, the number of startups declined by more than 25% from 2005 and 2013. More than 30% of startups fail before Year 2 and more than 50% of startups fail before Year 5. The rate of success for startups is 8% each year, meaning that more than 90% of all startups in any year either die or fail to hire more than 10 employees within 5 years.

[4] Source: Business Dynamic Statistics and Startup History File.

Characterizing the Startup Workforce

To characterize the workforce associated with each startup we create a Startup Worker History File (2005-2014) derived from individual level data on jobs. Universe data on jobs come from administrative records. Each paid job for each individual from 2005-2014 is reported at the Employer Identification Number (EIN) level via IRS form W2 and state-level Unemployment Insurance wage records. The latter underlie the core LEHD infrastructure (*25*) and are necessary to identify the establishment for the bulk of multi-unit firms (*30*). The combined data includes more than 2.6 billion person-EIN-year observations (approximately 1.83 billion match across the W2 and LEHD/UI universes, 550 million are found only in the W2 records and 320 million are only found in LEHD). We then enhance this data with the Individual Characteristics File (ICF), which includes demographic data on persons including sex, age, race and place of birth.[5] We are able to link 48 million of the 2.6 billion person-EIN-year observations to startups, giving us an average of nearly 4.5 million person-startup observations each year.[6]

We derive the human capital characteristics for each individual worker in the startup workforce at each time t from their work history in the previous three years. We create separate flags for whether the individual worked for (i) an R&D performing firm, (ii) a firm in a High Tech industry, (iii) a national research university and (iv) a national research university and paid on a research grant. The individual level data are then aggregated to create human capital composition measures for each startup for each year. The first three of these human capital measures are derived from a combination of different sources of internal Census Bureau data. The last is derived from new UMETRICS data combined with machine learning methods as described below.

The R&D measure is created from adding firm-identifiers based on the Business Innovation and Research and Development Survey (BRDIS) and Survey of Industrial Research and Development (SIRD)[7]. A firm is classified as an R&D firm if it has positive R&D expenditures during the year the employee was affiliated with the firm. The High Tech industry classification. The High-Tech industries is derived from work by Hecker (*31*, *32*), which is based on the relative concentration of STEM workers. The university measure is derived from data from IPEDS and the Carnegie Institute which provide a frame of universities in the United States. We then merge in national university research outlays collected by National Center for Science and Engineering Statistics at the National Science Foundation and keep the top 130 universities that comprise of 90% of total federally funded R&D research.

The identification of individuals working on research grants can be derived from UMETRICS data (*33*), which includes 14 universities accounting for 15% of federally funded research. The UMETRICS data are universe data from the personnel and financial records of universities. Although four files are provided by the university, the key file of interest in this project is the

[5] A detailed discussion on the matching process and match rates is provided in the appendix.
[6] This figure differs from the reported BDS statistics, which calculate employment at startups at a specific point in time (March 12). Our figures are higher, reflecting employee-employer transitions (i.e. workers who work briefly for a startup and then move to a different job). The 48 million observations represent 37.8 million unique individuals.
[7] We use the SIRD to identify R&D firms between 2005-2007 and BRDIS for 2008-2014

employee file. Briefly, for each funded research project, both federal and nonfederal, the file contains all payroll charges for all pay periods (identified by period start date and period end date) with links to both the federal award id (unique award number) and the internal university id number (recipient account number). In addition to first name and last name, and date of birth, the data include the employee's internal de-identified employee number, and the job title (which we mapped into broad occupational categories). Each university provided data as far back as they had reliable records (see Appendix for more details). We extend the measure to all universities and back to 2005 using machine learning approaches; that is discussed in the next section.

Machine Learning and Identifying Individuals funded from research grants

The current UMETRICS frame consists of 14 large research universities, with several concentrated in the Midwest. Although some have provided data from the early 2000s, the bulk provide data in the latter years of our sample. The current UMETRICS frame consists of 140,000 research trained individuals that can be linked to Census data and used to create a training dataset for machine learning purposes.

The training dataset consists of the employment and earnings records of all 14 UMETRICS universities in the period in which they provide data. By combining the UMETRICS and W2 data, we can identify all 140,000 who were employed on research grants in those time periods as well as 1.4 million who are not. The out-of-sample set includes 6.8 million individuals paid by the top 130 research universities in our time frame. Importantly, the out-of-sample set includes years for some UMETRICS institutions outside of those provided by the universities.

The link to Census data enables us to create a rich set of attributes that can be used to train the machine learning models. We are able to capture each employee's earnings history before, during and after the employee's time at the university, the dominant employer characteristics which include size, payroll, average earnings, industry, location and other-job earnings, in-state and out-of-state earnings, industry earnings, geographic variation (across all 50-states), university characteristics (collected from IPEDS, Carnegie Institute, NSF and NIH) which include average SAT scores, enrollment levels, public/private indicators, along with yearly variations and before/after/during (for the period t-2 until t+2 for the individual entering and exiting the university) across all variables. All of this is supplemented with demographic data collected from the Individual Characteristics File (ICF). In total, we have over 1,500 person-EIN level features to train the machine learning algorithms.

The success of our machine learning methods hinges on the extent to which there are measurable differences between research trained and non-research trained individuals. Table 1 below highlights some key differences between employees working on research grants and those not.

Table 1: Comparison of demographic and earnings characteristics

	Research Trained	Not Research trained
Proportion Female	50.5	54.1
Proportion White	73.2	77.2
Proportion Hispanic	4.3	4.9
Proportion Black	5.7	9.3
Proportion Asian	14.1	6.2
Proportion Foreign-Born	21.8	11.4
Year of Birth	1977.7	1975.6
Proportion in Professional/Scientific Services	18.4	14.3
Professional/Scientific Earnings, t+1	42,500	33,700

Source: W2 and UMETRICS data.
Note: Each of these are significantly different at $p<0.001$.

Research trained individuals tend to be disproportionately male, Asian, foreign-born and younger relative to non-research trained employees (employees at the same institution but not affiliated with research grants). Research trained individuals are also more likely to be employed in Professional and Scientific services subsequent to leaving the university and have an earnings premium that is 30% higher in Professional and Scientific services in the year immediately following their exit from the university.

The quality of our classification methods will also depend on the extent to which our UMETRICS universities are broadly representative of the 130 out-of-sample research universities. Table 2 compares the national university sample with the UMETRICS sample. As we can see, the majority of universities included in the sample are large, public universities with medical schools attached to them. The UMETRICS sample is slightly larger on average and expends more on R&D. There are approximately 6.8 million Out-of-Sample individuals employed at these universities between 2005 and 2014.

Table 2: Comparison of university characteristics

	130 Universities	UMETRICS Sample[8]
Mean R&D Expenditure ($000, 2014	424,600	661,700
Mean Non-R&D Expenditures ($000), 2014	20,400	35,800
Mean # of NIH Awards, 2014	270	440
Mean Annual Enrollment	30,800	43,400
Mean Amount of NIH Awards ($000), 2014	112,500	180,900
Mean Undergraduate Enrollment, 2014	19,800	27,700
Mean Bachelor Degrees Awarded, 2014	4,700	6,900
Mean Graduate Enrollment, 2014	7,900	11,800
Mean Master Degrees Awarded, 2014	1,900	3,100
Mean Doctoral Degrees Awarded, 2014	700	1,100
Mean Total Degrees Awarded, 2014	7,300	11,100
Mean Faculty Number, 2014	1,400	2,200
% Private	28.5	30.8
% Land Grant	40	61.5
% with Medical School	69.2	84.6
Mean SAT Combined, 2014	1,140	1,190

Source: IPEDS and the Carnegie Institute.

The objective of our machine learning approach is to classify individuals in the out-of-sample set as to whether or not they participated in (were paid by) grant funded research. Our methodology proceeds as follows. First, we execute several feature selection models. Second, we estimate a series of supervised learning classification models with different parameterizations. Third, we perform a number of cross validation exercises to assess the sensitivity and robustness of the in-sample predictions. Finally, we use our preferred specification to predict which of the 6.8 million out-of-sample individuals participated in grant-funded research.

We perform a series of feature selection exercises to reduce the number of attributes considered by each learning model. Feature selection can provide a number of benefits including avoiding over-fitting, reducing computational burden, and improving prediction quality by filtering low value added features and/or selecting a subset of the most valuable featured based on prediction quality(*34*). We explore several univariate feature selection methodologies including k-best chi squared and univariate k-best by decision tree precision. We also use mean decreased impurity in a multivariate random forest model(*35*). Finally, we develop some hand-curated feature sets based on iterative implementation and testing. Each of the resulting feature subsets are used to train the classification models.

For each of the k-best methods we select the top 50 features.[9] The k-best chi squared method estimates the chi-square test statistic between each feature and class (research training status) and selects the top *k* features based on those estimates. This method measures the dependence

[8]Ohio State University, Penn State, Purdue, Michigan State, New York University and the Universities of Arizona, Illinois (Champaign-Urbana), Iowa, Michigan, Missouri, Wisconsin.

[9] In the future, we plan to experiment with the 100 and 200 best by each method.

between each feature and class removing those that are most likely to be independent of research training status and therefore less useful for classification. The k-best decision tree method estimates a decision tree classifier for each feature and class individually and evaluates the quality of in-sample predictions based on that single feature. Intuitively, features that have less predictive value will produce lower quality predictions when used in a univariate classification model. Features are ranked according to the mean stratified 3-fold cross-validated precision score from fitting the decision tree classification model for each feature-class combination. Precision, discussed in more detail below, captures the probability that a randomly selected positive predicted research training status is true. For our purposes, precision is the most relevant measure since we are most interested in measuring economic outcomes associated with positively classified individuals in the out-of-sample set.

Multivariate feature selection methods improve upon univariate methods by incorporating the complex interactions that can occur between features in supervised learning classification models. We calculate the k-best features by mean decreased impurity (Gini importance) in a random forest classifier(*36*). The Gini importance measure is derived from the Gini index used to split the data at each node, which captures the level of impurity/inequality among samples assigned to a node based on the split from its parent node(*37*).

We estimate several classification models including logistic regression, decision tree, and random forest. The first classification model we estimate is a logistic regression classifier, a classic supervised learning method for binary classifications problems(*38*, *39*). This model serves as a baseline from which we compare the performance of the tree-based methods. The second classification model we estimate is the decision tree model(*36*). Finally, we estimate a series of random forest models with different parameterizations (*40*). We explore a series of evaluation metrics for in-sample predictions resulting from different parameterizations of the random forest classifier.

To evaluate our predictions, we calculate several quality measures including accuracy, precision, recall. We also use the share of false positives and false negatives to guide model selection and parameter tuning. Accuracy captures how often the model is correct with respect to both positive and negative classifications. This measure will tend to be less useful for our purposes since we are most concerned about correctly identifying positives (those that participated in grant funded research). Accuracy is defined in the following way.

$$ACCURACY = \frac{tp + tn}{tp + tn + fp + fn}$$

Where *tp*, *tn*, *fp*, and *fn* are true positives, true negatives, false positives, and false negatives respectively. Precision can be thought of as the probability that a randomly selected person predicted to have participated in grant funded research actually did. Recall, on the other hand, captures the probability that a randomly selected grant funded researcher was correctly classified. Since we are primarily interested in the quality of positive classifications, in the discussion that follows precision will be our primary measure of quality.

$$PRECISION = \frac{tp}{tp + fp}$$

$$RECALL = \frac{tp}{tp + fn}$$

The accuracy, precision, and recall measures estimated by training and predicting using the entire training set will suffer from over-fitting. To avoid this issue, and obtain a more accurate measure of model quality, we perform several cross validation exercises. First, we execute a stratified K-fold cross validation strategy. Second, we perform "Leave-One-Out" cross validation at the university level.

Using stratified K-fold cross validation, we segment the data into 10 folds stratified in such a way that each sample contains approximately the same relative frequency of observations within each class (research trained (1s) and non-research trained (0s)). We then cycle through each fold, training the classification algorithm using the K-1 samples and test on the Kth. For the "Leave-One-Out' cross validation we iterate over the UMETRICS universities leaving one out, training the model using the remaining universities and predict on the excluded university. This allows us to simulate the addition of a new university to the UMETRICS data.

Table 3 shows the in-sample evaluation metrics for the logistic regression and decision tree classification models using several feature selection sets.

Table 3: Logistic Regression and Decision Tree Classification Results

	Logistic Regression			Decision Tree		
Feature Set	Chi-Squared	Decision Tree	Impurity	Chi-Squared	Decision Tree	Impurity
In-Sample Accuracy	88.405	88.431	88.422	99.984	97.847	99.996
In-Sample Precision	32.090	30.000	33.566	99.991	99.254	99.995
In-Sample Recall	0.274	0.064	0.170	99.872	81.987	99.970
Mean 10-Fold Precision	30.034	36.656	36.852	28.158	27.386	31.542

Source: UMETRICS, W2, LEHD, LBD, ICF and BR.

The results in Table 3 show that while accuracy is relatively high with the logistic regression classifier, it generally fails to predict research trained individuals with precision of roughly 31 across the different feature sets. Moreover, the recall for the logistic regression model is very poor. The decision tree results for all three feature sets, on the other hand, appear very promising with nearly perfect accuracy and precision. However, in the stratified 10-fold validation we see that while the logistic regression model retains its precision scores of roughly 30 in the cross validation, the decision tree model performs significantly worse in cross validation. This suggests that the decision tree tends to over-fit the training sample. Table 4 shows the results from the random forest classifier across the feature sets.

Table 4: Random Forest Classification Results

	Random Forest			
Feature Set	Chi-Squared	Decision Tree	Impurity	Hand-Curated
Estimators	50	50	50	50
Maximum Depth	50	50	50	50
In-Sample Accuracy	99.703	97.277	99.950	99.924
In-Sample Precision	99.990	98.971	99.991	99.981
In-Sample Recall	97.443	77.248	99.580	99.365
Mean 10-Fold Precision	99.849	70.816	99.806	62.222
Mean University-Fold Precision	86.873	86.661	86.710	

Source: UMETRICS, W2, LEHD, LBD, ICF and BR.
Note: Fewer than 50 estimators and lower maximum depth results in significant loss of precision and accuracy while additional estimators and depth yield little additional quality improvements and entail significant additional computational resources. The hand-curated set includes demographic variables and demeaned earnings for individuals during their time at the university.

The results in Table 4 suggest that the random forest classifier, by aggregating many different decision trees, avoids some of the over-fitting issues in the decision tree results. The accuracy, precision, and recall across the different features sets are high, with exception of the recall score for the univariate decision tree feature set, which drops from over 97 to about 77. This pattern is also evident in Table 3, where we see the univariate decision tree produces lower recall scores for both the logistic regression and decision tree classifiers. We also show in Table 4 the results using a hand-curated set of features, which includes demographic variables and demeaned earnings. We create this hand curated set by iteratively experimenting with different combinations of features to balance the quality of in-sample predictions with the number of out-of-sample positive predictions.

Applying our preferred classification model, the random forest estimator with the hand-curated feature set, to the out-of-sample set identifiers an additional 188,000 individuals who are likely to be research trained. Table 5 below compares the out-of-sample results with the in-sample and individuals not likely to be research-trained.

Table 5: Comparison of Economic and Demographic Characteristics

	Research Trained		Not research trained
	In Sample	Out of Sample	
Proportion Female	50.5	47.8	54.1
Proportion White	73.2	67.6	77.2
Proportion Hispanic	4.3	7.1	4.9
Proportion Black	5.7	5.6	9.3
Proportion Asian	14.1	16.2	6.2
Proportion Foreign-Born	21.8	25.6	11.4
Year of Birth	1977.7	1976.2	1975.6
Proportion in Professional/Scientific Services	18.4	18.4	14.3
Professional/Scientific Earnings, t+1	42,500	41,250	33,700

Note: Each of the differences listed in this table are statistically significantly different at $p<0.001$.

The out-of-sample prediction of research training compares favorably with the known in-sample group of research trained individuals. There are a couple of notable differences however. The out-of-sample is significantly more likely to be male, Hispanic and foreign-born than the in-sample.

Our combined data consists of a national sample of Startups and their outcomes between the years 2005 and 2014, as well as a national sample of all workers affiliated with these startups, along with 4- main designations of human capital attributes assigned to each worker. The next section explores some basic summary tables and findings for these different types of workers and their potential impact on startups.

Basic Facts

This section establishes some basic facts on the human capital composition of the startups by year, as well as startup outcomes. As Figure 1 shows, the majority of Startups fail within 5-years and more than 90% of Startups either die or hire fewer than 10 employees within the first 5-years of existence. Of course, not all startups are the same and since we are primarily interested in the dynamism of the US economy, we will also focus on high growth startups within sectors likely to employ a research trained workforce including “industrial” startups (defined as being a startup engaged in either manufacturing, information technology, finance, professional/scientific services and health care), and High Tech startups (classified according to STEM concentration).

Figure 2 shows the size distribution for all startups, along with their average earnings distribution at time $t=0$.

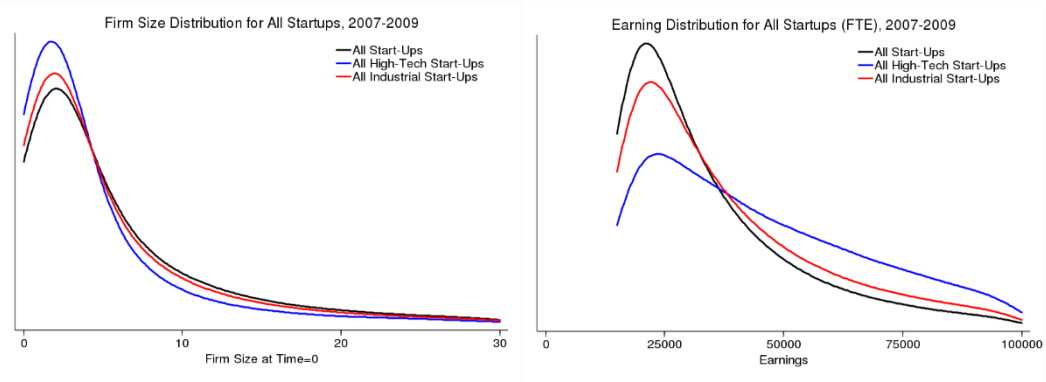

Figure 2: Startup Size and Earnings Distribution at time *t=0*

The vast majority of Startups are extremely small in their first year as 75% of all startups have fewer than 5 employees at time *t=0*, with more than 50% of startups having 2 or fewer employees. Fewer than 5% of Startups hire more than 20 employees in the initial period. This is consistent across all startup types as well. Also, most Startups offer relatively small earnings, with startups in High Tech industries typically offering the highest earnings. These two findings, combined with the high-rate of failure suggest that startups face significant capital. The small size also highlights the importance of human capital in the initial period.

Human capital composition

Table 6 below provides the total number of startup employees, along with the proportion of employees that have R&D experience, High-Tech experience, University experience and research grant experience[10] within the 3-years prior to joining the startup.

Table 6: Startup Employment Composition[11]

Year	Total ever employed at startup	R&D Experience	High Tech Experience	University Experience	Research experienced
2006	6.82M				0.09
2007	6.47M				0.09
2008	5.74M				0.09
2009	4.7M	19.3	11.1	2.6	0.09
2010	4.56M	20.3	12	2.2	0.10
2011	4.37M	21.2	13.7	2.4	0.10
2012	4.53M	21.1	13.4	2.6	0.09
2013	4.4M	22.2	14.2	2.7	0.09

[10] Note that about 25,000 of the research experienced individuals working in startups are directly identified through UMETRICS data. The balance are derived from the machine learning algorithm

[11] Since we focus on the prior 3-years work experience, the table is left-censored

Source: Startup Worker History and Startup Firm History Files.

Approximately one in five workers in a startup has experience in an R&D performing firm and one in ten has experience in a High Tech firm. About 3% of the startup workforce is affiliated with a university in the 3-years prior to the startup, and roughly 5% of the university affiliated workforce has worked on a research grant.

Table 7 shows the human capital composition by startup type.

Table 7: Human-Capital Composition by Startup Type

	Former High-Tech Employees	Former R&D Employees	Former University Employees
All Startups	10%	17%	2%
High Tech Startups	94%	26%	4%
Industrial Startups	16%	20%	3%

Source: Startup Worker History and Startup Firm History Files.

The table makes it clear that High Tech startups are nearly entirely composed of High Tech employees and have much greater proportions of workers who were previously at R&D performing firms. They also have twice as many former university employees as other startups. Similarly, industrial startups have higher proportions of employees with experience at High Tech and R&D performing firms, as well as more employees with university experience.

Startup Outcomes and workforce characteristics

This section provides some initial descriptive results about the link between workforce experience and startup outcomes.

The outcome variables of interest are measured as follows:

1. Survival to period *t+1*,
2. Success (defined as having survived for at least 5-years and employ 10+ employees at time *t+5*),
3. High Growth (defined as having survived for at least 5-years, employ 10+ employees at time *t+5* and be in top ten percentile of employment growth among your cohort (conditional on employing 5+ employees at time *t=0*)),
4. Employment Growth to *t+1* (conditional on having at least 5+ employees at time *t=0*),
5. Employment Growth to *t+5* (conditional on having at least 5+ employees at time *t=0*).

We standardize our descriptive analysis by defining a startup's workforce as "intensive" in one of our human capital dimensions if it employs more of a certain type of worker than the median startup within a size group. This means that for all startups of size 10 at time t=0 for example, we compare the outcomes of startups that employ disproportionately more R&D workers to startups

that employ disproportionately less R&D workers. The results for survival outcomes are reported below in Figure 3.

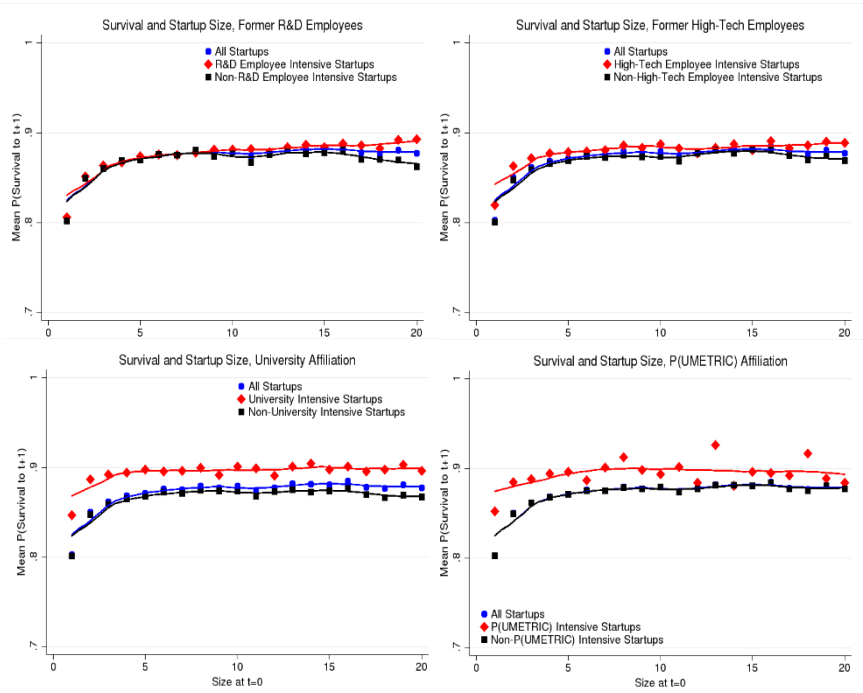

Figure 3: Survival by Human capital intensity

Figure 3 is consistent with the view that startups with higher proportions of high human capital employees are more likely to survive. We see a clear separation in the survival probabilities of startups that hire University employees and research trained (UMETRICS) employees intensely. There is minor separation in the survival probabilities for High Tech startups and almost no difference in the survival probabilities between employees with and without experience in R&D performing firms.

Figure 4 shows the results of a similar analysis using a measure of whether the startup was successful (defined as having 10+ employees and surviving for 5+ years).

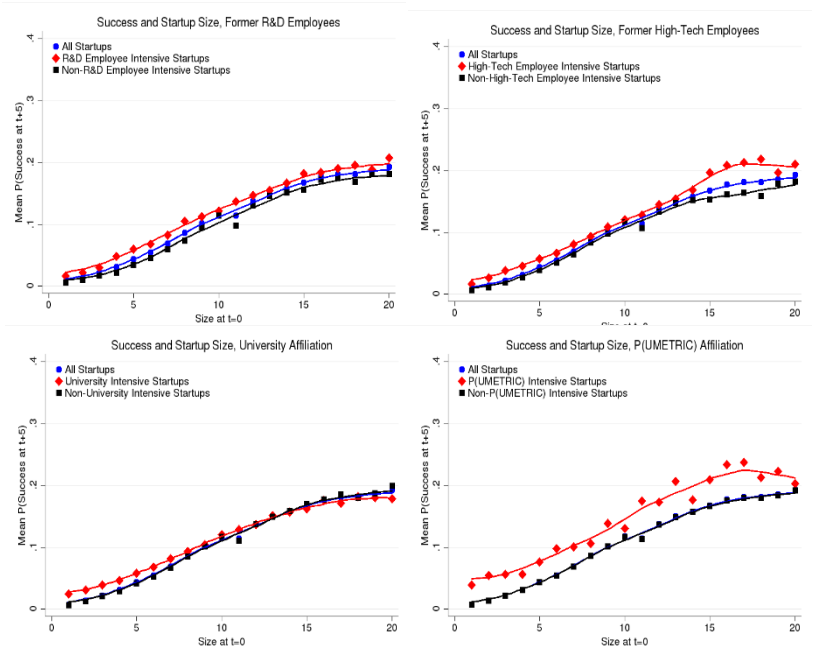

Figure 4: Startup Success and Human Capital Intensity

There are minor differences in the probability of startup success for those startups that hire R&D employees intensively, as well as startups that hire High Tech employees intensively. Interestingly, there is almost no difference in the success outcomes for startups that hire university employees intensively. However, there is a substantial difference in the success outcomes for startups that hire research trained (UMETRICS) employees.

Finally, Figure 5 shows the results of a similar exercise that compares the outcomes for whether a startup is a high growth startup (defined as having 10+ employees and being in the top 10% of the employment growth rate distribution within their startup year cohort).

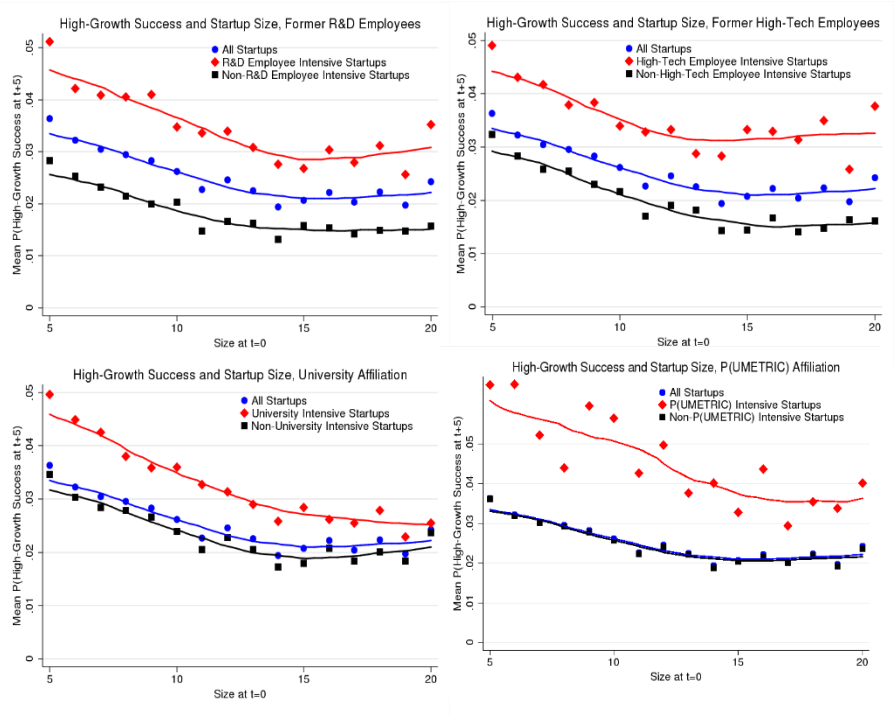

Figure 5: High-Growth Success and Human Capital Intensity

There are clear differences in the probability of high growth success across all designations of human capital. The results are consistent with those shown in Figures 3 and 4, but do display more dispersion. This may be due to the fact that these high growth firms make up fewer than 1% of the total number of startups, creating more volatility and disclosure restrictions for the subset of firms that hire more than 20 employees in the initial period.

Analytical Results

The basic framework was provided in Equation (1). We assume that the functional form of Equation (1) is a linear combination of exponential functions, allowing us to use a log-linear estimation and calculate multiple outcome measures for each startup (survival, "success", "high-growth success" and employment growth) both one and five years after the birth of the firm. We regress these outcomes against the startup's workforce and other characteristics in the year of firm birth (t=0).

Our main empirical specification is as follows

$$Y_f = \alpha + \beta_1 \ln EARN_{f0} + \sum_{k=1}^{9} \delta_k SIZE_{kf0} + \beta_2 \ln \overline{AGE_{f0}} + \beta_3 \ln FEMALE_{f0} + \beta_4 \ln FOREIGN_{f0} + \beta_5 \ln RD_{f0} + \beta_6 \ln HT_{f0} + \beta_7 \ln UNI_{f0} + \beta_8 \ln Research\ Experience_{f0} + \varepsilon$$

The key measures of interest are the workforce human capital measures – the number of workers who have worked in R&D performing firms, High Tech firms, universities – as well as the number who have direct research experience. Since the Census Bureau does not have direct measures of technology, we control for industry, detailed geography and year. We also include mean earnings of the workforce as well as firm employment size categories. External macroeconomic conditions are proxied by zip code-year fixed effects and industry fixed effects.

The first specification separates all of the human capital designations in order to separately describe the relationship between each type of human capital and startup outcomes before applying control factors.

Figure 6 reports the coefficient estimates of the standalone human capital designations by firm-size for 2 separate outcomes: survival and success. The results show that the standalone human capital coefficients decline as the firm gets bigger, highlighting that there may be diminishing marginal returns to the employment of each additional type of worker. The returns to each type of worker declines very rapidly for the survival outcome, with a more modest and steady decline in the success rates.

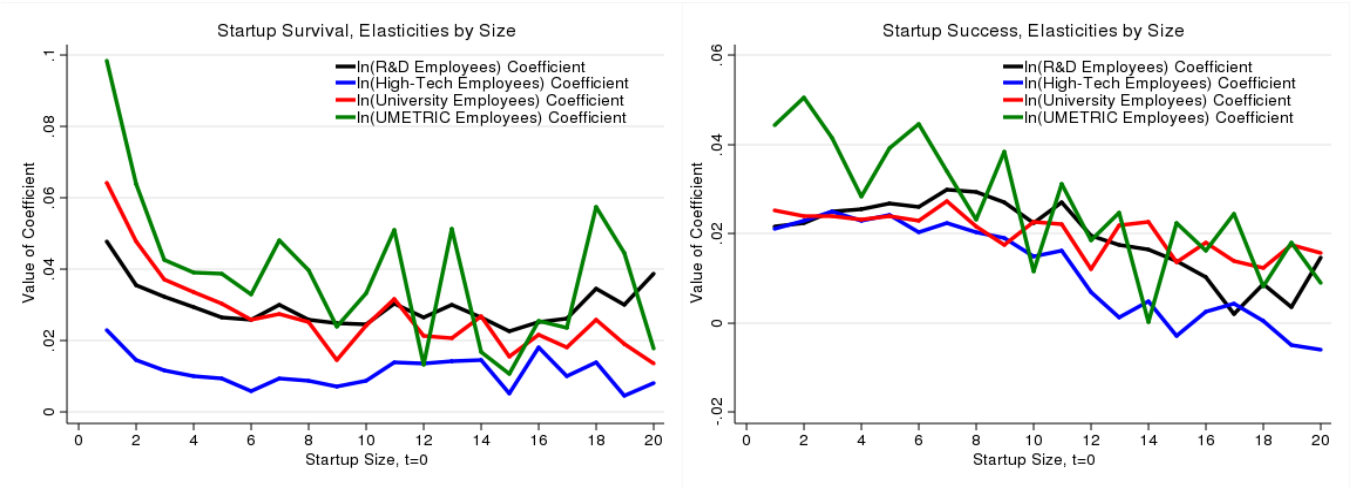

Figure 6: Coefficient Values of Standalone Human Capital Measures by Firm Size

Table 8 provides the key results associated with the full regression, including all control variables. Briefly, all measures of workforce R&D experience are positively and significantly related with startup survival and success. A one unit increase in the worker-type leads to approximately 0.07%-2.45% increase in the survival rate to year t+1, a 1.6%-7.7% increase in the probability of becoming a high-growth success and between 2-6% increase in the employment growth rate. The effects are less dramatic depending on the human capital measure, but each measure shows a consistent positive relationship. Interestingly, the strongest effect appears to be for employment growth both one and five years from birth. The results are very similar regardless of whether interaction terms are or not included, or whether startups are classified by starting size (see Appendix for details).

Table 8: OLS on All Startup Outcomes, 2005-2014

Outcome Variable	Survival, year 1	Success, year 5	High Growth Success, year 5	Employment Growth, year 1	Employment Growth, year 5
$\ln RD_{f0}$	0.0245***	0.0149***	0.00564***	0.195***	0.177***
	(0.000528)	(0.000306)	(0.000135)	(0.00146)	(0.00335)
$\ln HT_{f0}$	0.000708	0.00747***	0.00444***	0.0586***	0.0784***
	(0.000685)	(0.000396)	(0.000175)	(0.00179)	(0.00422)
$\ln UNI_{f0}$	0.0130***	-0.00347***	0.00160***	0.0570***	0.0550***
	(0.00112)	(0.000648)	(0.000287)	(0.00266)	(0.00649)
$\ln research\ experience_{f0}$	0.0169***	0.0285***	0.00766***	0.00841	0.0245
	(0.00394)	(0.00228)	(0.00101)	(0.00832)	(0.0189)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Observations	3,730,000	3,730,000	3,730,000	757,000	259,000
R-squared	0.110	0.151	0.034	0.209	0.148

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001; controls included for size and average earnings, proportion of workforce that is female, foreign born, and interactions of female, foreign born with research experience. Full results in the appendix

Tables 9 and 10 report the results for two different categories of startups - industrial startups and high-tech startups. The results are substantively unchanged but there are a few noticeable differences.

Table 9: OLS on Industrial Startup Outcomes, 2005-2014

Outcome Variable	Survival, year 1	Success, year 5	High-Growth Success, year 5	Employment Growth, year 1	Employment Growth, year 5
ln RD_{f0}	-0.00611***	0.0151***	0.00771***	0.195***	0.201***
	(0.00105)	(0.000649)	(0.000315)	(0.00317)	(0.00744)
ln HT_{f0}	0.0162***	0.0190***	0.00573***	0.142***	0.182***
	(0.00104)	(0.000648)	(0.000314)	(0.00341)	(0.00803)
ln UNI_{f0}	0.00908***	-0.000379	0.00144*	0.0787***	0.0908***
	(0.00186)	(0.00115)	(0.000559)	(0.00510)	(0.0129)
ln $research\ experience_{f0}$	0.00476	0.0100*	0.00803***	-0.0372*	0.0365
	(0.00683)	(0.00423)	(0.00206)	(0.0170)	(0.0411)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Observations	1,134,000	1,134,000	1,134,000	194,000	73,000
R-squared	0.107	0.164	0.052	0.303	0.205

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001; controls included for size and average earnings, proportion of workforce that is female, foreign born, and interactions of female, foreign born with research experience. Full results in the appendix

We find that employing workers that worked previously for R&D firms has a negative and significant impact on survival for "industrial" startups, but an extremely strong impact on employment growth rates (especially, relative to the other human capital measures). This is consistent with the idea that hiring of these worker-types is especially risky. The impact of high-tech workers however is positive and significant across all outcomes and especially high. The impact of university-affiliated employees on industrial startups is also positive, while the impact of university-affiliated workers with research experience is modest, but very high in terms of yielding high-growth successful startups.

Turning now to high-tech startups, we have Table 10.

Table 10: OLS on High Tech Startup Outcomes, 2005-2014

Outcome Variable	Survival, t+1	Success, t+5	High-Growth Success, t+5	Employment Growth, t+1	Employment Growth, t+5
ln RD_{f0}	-0.0436***	0.00979***	0.00748***	0.0558***	0.146***
	(0.00262)	(0.00153)	(0.000776)	(0.0101)	(0.0269)
ln HT_{f0}	0.0827***	0.0493***	0.00992***	0.531***	0.488***
	(0.00247)	(0.00145)	(0.000733)	(0.0126)	(0.0342)
ln UNI_{f0}	-0.00359	-0.0178***	-0.00591***	0.0494***	-0.0568
	(0.00435)	(0.00254)	(0.00129)	(0.0146)	(0.0439)
ln $research\ experience_{f0}$	0.00765	0.0154	0.0122**	-0.0325	0.0140
	(0.0152)	(0.00885)	(0.00450)	(0.0443)	(0.123)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Observations	148,000	148,000	148,000	17,000	6,400
R-squared	0.144	0.168	0.095	0.522	0.427

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001; controls included for size and average earnings, proportion of workforce that is female, foreign born, and interactions of female, foreign born with research experience. Full results in the appendix

We see a mostly similar ordering of human capital coefficients as the previous table, with the exception of the high-tech worker coefficients, which are the key drivers of employment growth, survival and success for high-tech startups. Given the plethora of anecdotal evidence on the reliance of high-tech startups on previous high-tech experience, the importance of hiring high-tech workers in determining outcomes makes plenty of sense.

Summary

This paper leverages new data about workforce human capital that can be used to provide more insights into the survival and employment growth of new businesses. These results are consistent with the view that there is a relationship between workforce experience and business startup and survival. Further work using these data will be necessary to examine temporal dynamics. It will be particularly interesting to understand whether changes in the fluidity of this type of workforce, or changes in the nature of research funding, can be tied to the decline in business dynamism.

References

1. I. Hathaway, R. E. Litan, Declining business dynamism in the United States: A look at states and metros. *Brookings Inst.* (2014).
2. C. I. Jones, Sources of US economic growth in a world of ideas. *Am. Econ. Rev.* **92**, 220–239 (2002).
3. D. Acemoglu, U. Akcigit, N. Bloom, W. R. Kerr, "Innovation, reallocation and growth" (National Bureau of Economic Research, 2013).
4. N. Bania, R. W. Eberts, M. S. Fogarty, Universities and the startup of new companies: can we generalize from route 128 and Silicon valley? *Rev. Econ. Stat.*, 761–766 (1993).
5. R. A. Lowe, C. Gonzalez-Brambila, Faculty entrepreneurs and research productivity. *J. Technol. Transf.* **32**, 173–194 (2007).
6. N. Hausman, "University Innovation, Local Economic Growth, and Entrepreneurship" (2012), (available at http://ideas.repec.org/p/cen/wpaper/12-10.html).
7. E. L. Glaeser, W. R. Kerr, G. A. M. Ponzetto, Clusters of entrepreneurship. *J. Urban Econ.* **67**, 150–168 (2010).
8. S. Kantor, A. Whalley, Research proximity and productivity: long-term evidence from agriculture. *Rev. Econ. Stat. Forthcom.* (2014).
9. S. Kantor, A. Whalley, Knowledge Spillovers from Research Universities: Evidence from Endowment Value Shocks. *Rev. Econ. Stat.* **96**, 171–188 (2013).
10. L. Fleming, I. I. I. Charles King, A. Juda, Small Worlds and Regional Innovation. *Organ. Sci.* **18**, 938–954 (2007).
11. M. Marx, J. Singh, L. Fleming, Regional disadvantage? Employee non-compete agreements and brain drain. *Res. Policy*. **44**, 394–404 (2015).
12. R. Decker, J. Haltiwanger, R. Jarmin, J. Miranda, The role of entrepreneurship in US job creation and economic dynamism. *J. Econ. Perspect.*, 3–24 (2014).
13. W. R. Kerr, R. Nanda, M. Rhodes-Kropf, "Entrepreneurship as experimentation" (National Bureau of Economic Research, 2014).
14. T. Åstebro, H. Herz, R. Nanda, R. A. Weber, Seeking the roots of entrepreneurship: insights from behavioral economics. *J. Econ. Perspect.* **28**, 49–69 (2014).
15. R. A. Decker, J. Haltiwanger, R. S. Jarmin, J. Miranda, Where has all the skewness gone? The decline in high-growth (young) firms in the US. *Eur. Econ. Rev.* **86**, 4–23 (2016).
16. F. Karahan, B. Pugsley, A. Sahin, Understanding the 30-year decline in the startup rate: A general equilibrium approach. *Unpubl. manuscript, May* (2015).
17. A. Tabarrok, N. Goldschlag, 9. Is Entrepreneurship in Decline? *Underst. Growth Slowdown*, 169 (2015).
18. R. J. Gordon, *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War* (2016), vol. 1.
19. N. Bloom, C. I. Jones, J. Van Reenen, M. Webb, "Are Ideas Getting Harder to Find" (2016).
20. C. Syverson, "What determines productivity?" (National Bureau of Economic Research, 2010).
21. R. Molloy, C. L. Smith, R. Trezzi, A. Wozniak, Understanding declining fluidity in the US labor market (2016).
22. A. Golan, J. Lane, E. McEntarfer, The dynamics of worker reallocation within and across industries. *Economica*. **74**, 1–20 (2007).
23. M. Bjelland, B. Fallick, J. Haltiwanger, E. McEntarfer, Employer-to-employer flows in

the united states: estimates using linked employer-employee data. *J. Bus. Econ. Stat.* **29**, 493–505 (2011).
24. S. P. Kerr, W. Kerr, Ç. Özden, C. Parsons, Global talent flows. *J. Econ. Perspect.* **30**, 83–106 (2016).
25. J. M. Abowd, J. Haltiwanger, J. Lane, Integrated Longitudinal Employer-Employee Data for the United States. *Am. Econ. Rev.* **94**, 224–229 (2004).
26. E. Barth, J. Davis, R. B. Freeman, Augmenting the Human Capital Earnings Equation with Measures of Where People Work. *Natl. Bur. Econ. Res. Work. Pap. Ser.* **No. 22512** (2016), doi:10.3386/w22512.
27. J. M. Abowd *et al.*, in *Measuring capital in the new economy* (University of Chicago Press, 2005), pp. 153–204.
28. J. Lane, J. Owen-Smith, R. Rosen, B. Weinberg, "New linked data on science investments, the scientific workforce and the economic and scientific results of science" (2014).
29. N. Zolas *et al.*, Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. *Science (80-.).* **350**, 1367–1371 (2015).
30. J. M. Abowd *et al.*, in *Producer Dynamics: New Evidence from Micro Data* (University of Chicago Press, 2009), pp. 149–230.
31. D. E. Hecker, High-technology employment: a NAICS-based update. *Mon. Lab. Rev.* **128**, 57 (2005).
32. N. Goldschlag, J. Miranda, Business Dynamics Statistics of High Tech Industries (2016).
33. J. I. Lane, J. Owen-Smith, R. F. Rosen, B. A. Weinberg, New linked data on research investments: Scientific workforce, productivity, and public value. *Res. Policy* (2015), doi:10.1016/j.respol.2014.12.013.
34. I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
35. R. Kohavi, G. H. John, Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997).
36. L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and regression trees* (CRC press, 1984).
37. C. Zhang, Y. Ma, *Ensemble machine learning* (Springer, 2012).
38. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
39. G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning (Vol. 103) (2013).
40. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).