I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11738

Occupational Classifications: A Machine Learning Approach

Akina Ikudo
Julia Lane
Joseph Staudt
Bruce Weinberg

AUGUST 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11738

Occupational Classifications: A Machine Learning Approach

Akina Ikudo
University of California, Los Angeles

Julia Lane
New York University, U.S. Census Bureau and IZA

Joseph Staudt
U.S. Census Bureau

Bruce Weinberg
Ohio State University and IZA

AUGUST 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Occupational Classifications: A Machine Learning Approach[1]

Characterizing the work that people do on their jobs is a longstanding and core issue in labor economics. Traditionally, classification has been done manually. If it were possible to combine new computational tools and administrative wage records to generate an automated crosswalk between job titles and occupations, millions of dollars could be saved in labor costs, data processing could be sped up, data could become more consistent, and it might be possible to generate, without a lag, current information about the changing occupational composition of the labor market. This paper examines the potential to assign occupations to job titles contained in administrative data using automated, machine-learning approaches. We use a new extraordinarily rich and detailed set of data on transactional HR records of large firms (universities) in a relatively narrowly defined industry (public institutions of higher education) to identify the potential for machine-learning approaches to classify occupations.

JEL Classification: J0, J21, J24

Keywords: UMETRICS, occupational classifications, machine learning, administrative data, transaction data

Corresponding author:
Julia Lane
Wagner School of Public Policy
New York University
The Puck Building
295 Lafayette Street
New York, NY 10012-9604
USA

E-mail: julia.lane@nyu.edu

[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. Lane was supported through an Intergovernment Personnel Act assignment to the US Census Bureau. The research agenda draws on work with many coauthors, but particularly Jason Owen Smith.

NON-TECHNICAL SUMMARY

We use a new source of data to examine the potential to use automated techniques to generate occupational categories. If it were possible to combine new computational tools and administrative wage records to generate an automated crosswalk between job titles and occupations, millions of dollars could be saved in labor costs, data processing could be sped up, data could become more consistent, and it might be possible to generate, without a lag, current information about the changing occupational composition of the labor market.

While our results suggest that occupations can be assigned from job titles, they also point to real challenges. In particular, our analysis suggests that there are substantial limits to using machine learning to create discrete occupational categories, even with rich data sources. Our experience suggests that, rather than trying to generate occupational categories, it might be more sensible to directly generate information about the tasks performed on jobs and the skills and experience required by each job, especially given the increased emphasis on tasks in the literature in labor economics.

I. Introduction

Characterizing the work that people do on their jobs is a long-standing and core issue in labor economics. Traditionally, classification has been done manually, but there is a long literature on the associated challenges, well summarized by an influential paper by Mellow and Sider (*1*). Many organizations, including the Census Bureau and the Bureau of Labor Statistics, are beginning to investigate the potential of using new computational tools, such as text analysis and machine learning, to automatically classify workers' occupations (*2*). At the same time, there has been a surge of interest in using administrative wage records to directly capture occupations to inform training(*3*), to permit deeper longitudinal analysis on career outcomes, the effects of training, and changes in inequality. The problem is that standardized occupations do not exist on wage records, since they are drawn from human resource files which include firm specific job titles. If it were possible to combine new computational tools and administrative wage records to generate an automated crosswalk between job titles and occupations, millions of dollars could be saved in labor costs, data processing could be sped up, data could become more consistent, and it might be possible to generate, without a lag, current information about the changing occupational composition of the labor market.

This paper examines the potential to assign occupations to job titles contained in administrative data using automated, machine-learning approaches. Although there has been little research that directly ties firm-level human resource (HR) data on job titles to occupational classifications, there are intellectual foundations for occupational coding that are largely grounded in the survey world. The first foundation is conceptual: to define each occupation. The second is operational: to translate concepts to standardized protocols. The third is statistical: to infer occupations from the information at hand. The fourth pertains to resources: the implementation of such classifications at massive scale given the limited resources available. More generally, we contribute to a much larger set of classification problems, which are increasing in salience with the availability of more transaction data. It is important to understand which tools and approaches enable the

new, rich, but unstructured data to be used, while minimizing the need for expensive and slow manual classification.

We use a new extraordinarily rich and detailed set of data on transactional HR records of large firms (universities) in a relatively narrowly defined industry (public institutions of higher education) to identify the potential for machine-learning approaches to classify occupations. This is, to our knowledge, the first large-scale dataset that draws from such HR records across multiple institutions. These data have several advantages. First, the institutions are relatively large and complex, and they use HR systems that are similar to other large and complex organizations in the rest of the economy. Second, the focus on one industry limits the number of possible occupational categories, permitting a targeted analysis. Third, the focus on public universities is attractive because the HR descriptions associated with job titles are available online, and can be used to provide additional information for classification purposes. Finally, the industry is interesting in its own right. Indeed, the production of research often involves the use of intangible assets, particularly, labor inputs. Moreover, the training that students and postdocs receive is economically valuable (Zolas et al. 2015).

We build a training dataset from the HR records using human curation and additional rich data sources. First, university staff and trained students manually assign occupations to job titles. That manual curation is then enhanced with additional information from online job descriptions as well as Census Bureau micro-level information on demographic characteristics and earnings. The data are then used to train machine learning models to predict occupations from job titles. Finally, the results are evaluated.

While our results suggest that occupations can be assigned from job titles, they also point to real challenges. In particular, our analysis suggests that there are substantial limits to using machine learning to create discrete occupational categories, even with rich data sources. There are two core problems. The first is that occupational classifications are inherently noisy, so it is difficult to identify ground truth, particularly in a dynamic and changing economy. The second is that job titles have insufficient consistency or detail across institutions necessary for robust supervised machine learning. We do find that a large number of relatively sparsely populated job titles – a quarter of titles have only one

employee and over half have fewer than ten employees – could be assigned algorithmically, greatly reducing cost with little impact on accuracy. However, our experience suggests that, rather than trying to generate occupational categories, it might be more sensible to directly generate information about the tasks performed on jobs and the skills and experience required by each job, especially given the increased emphasis on tasks in the literature in labor economics (*5*).

The rest of the paper is organized as follows. Section II provides background. Section III describes data and our framework. Section IV and V provide detailed description of the first two principles, conceptual and operational, followed by statistical and implementation described in Section VI. Section VII concludes.

II. Background

One reason for developing occupational classifications is deeply rooted in sociology (*6*), as intrinsic to the measurement of the sources of inequality, social stratification and class mobility. Occupational classification is also essential in economic analyses, describing structural changes caused by technological advancement, automation, globalization, and change in immigration laws (*7*). Another reason for developing occupational classification is to provide an easy-to-measure pathway from generally understood job activities to skill needs in the economy (*8*).

In practice, occupational categories were developed at scale in the 1930's and codified in the Dictionary of Occupational Titles (*9*). Occupational analysts would "interview and observe workers, and then write job descriptions and make ratings of the characteristics of the occupation. To illustrate the magnitude of these efforts, over 75,000 on-site job analyses were conducted by analysts in the US Employment Services field centers throughout the country between the mid 1960's and 1970's alone" (Peterson et al. 2001, p. 453). Since then, the Department of Labor developed O*NET, which provides detailed occupational taxonomies and a detailed occupation-to-skills crosswalk (*9*).[2]

---

[2] The framework for developing occupational categories is based on (i) delineating the content domain of the occupations, (ii) developing common descriptors for assessment and (iii) developing rules for creating categories (*25*). It is worth noting that of the key principles adopted for developing taxonomies, the

The current approach to occupational classifications is thorough and thoughtful, but quite costly. The Office of Management and Budget has established a Standard Occupational Classification Policy Committee which is charged with both developing a uniform classification system and updating it on a regular basis (*10*). The process is extensive and time consuming. For example, the 2010 updating exercise involved establishing six working groups, and publishing a Federal Register request in 2006 and asking the public to comment on the classification principles used for the 2000 Standard Occupational Classification (SOC) system as well as corrections to the 2000 SOC manual, and suggestions for new categories (*11*, *12*). The 2018 updating exercise began in early 2012, a Federal Register notice was published in 2014, responses were provided in 2016 and the results were published at the end of 2017.[3]

In addition to cost, the measurement challenges with categorizing worker occupations on surveys are well known: they are notoriously noisy (*13*). In probably the best known analysis, Mellow and Sider find that only 83.3% of CPS respondents' major (1 digit) occupations match their employer's reports and that share falls to 59.7% for detailed (3 digit) occupations (and these rates are considerably lower than those for industry of employment, at 93.1% and 85.4% for major and detailed industry) (*1*). Bound et al. find similar errors in their overview of measurement errors (*14*), as do Abraham and Spletzer (*15*). Fisher and Houseworth find that there is systematic inflation of occupations for lower-skilled individuals (*16*).

Despite these measurement issues, there have been repeated calls to require firms to report occupational data as part of their federal reporting requirements. Indeed, a recent Department of Labor study group rated this need highest of all possible reporting activities (*17*). The cost of doing so manually might well be prohibitive – the state of Texas surveyed businesses and estimated "that the initial cost to employers could range from $478 million to $1.2 billion, with annual recurring costs of $342 million to $715

---

principles of exclusivity and exhaustivity means that the classification system is based on tasks, or work, performed, rather than skills or credentials. The reasoning is that since most occupations require multiple skills, it would be impossible to exclusively classify occupations if they were skill based (*11*, *12*).

[3] https://www.bls.gov/soc/2018/home.htm

million. Costs to the Texas Work Commission were estimated at $3.1 million in the first year, and a total five-year cost of $7.9 million to collect this data". (Texas Workforce Commission 2016, p. 17).

There have been some attempts to incorporate machine-learning methods into occupational classifications from open-ended survey questions (*19*, *20*). This approach is very different from ours, since we use administrative information on job titles, rather than survey responses. That work found that automated coding was feasible if there is sufficient training data. It emphasized the importance of data preprocessing, algorithmic quality – extending naïve Bayesian approaches to random forests – and thoughtful use of distance metrics in improving occupational prediction. It also suggested that machine learning might also have value by providing responders with candidate occupations as part of a learned cluster, rather than as part of a constructed and hierarchical decision tree.

III. Data and Framework

The administrative data we use are derived from the UMETRICS project, which builds on and extends the federal STAR METRICS effort (*21*). These data are maintained by the Institute for Research on Innovation and Science (IRIS) at the University of Michigan and currently contain record-level information on all wage payments made to individuals through research grants at 26 participating research universities (*21*, *22*). 62 university campuses have committed to join IRIS, accounting for just under half of federal university R&D expenditures and the projected membership in the next three years will include 90% of university research funding. In the interest of homogeneity, for our analysis, we chose large public research universities in the Big 10.[4]

Although multiple files are provided by the universities, we focus on the employee file, which for each federally funded project, contains all payroll charges for all pay periods

[4] The universities are Indiana, Wisconsin, Iowa, Michigan, Minnesota, Penn State, Rutgers and Ohio State University.

(period start date to period end date) with links to both the federal award ID (unique award number) and the internal university ID number (recipient account number). Also available from the payroll records are the employee's internal de-identified employee number, the job title, their FTE status and the proportion of earnings allocated to the award. In addition, the UMETRICS program has incorporated additional fields (notably, the name and date of birth of those supported on federally funded projects) to enable data linkage, and has enhanced the core data with additional information on grants derived from public sources.

We view these data as a valuable laboratory for quantifying the prospects for a machine-learning approach to occupation classification. In some ways these universities are well-suited to a machine-learning approach – they are large, generally similar, and highly structured. Thus, we can identify many different categories of workers and assess our ability to identify similar workers at other institutions. On the other hand, the uniformity of these institutions makes our task somewhat more challenging in that we need to make relatively fine distinctions (e.g. a dataset that is comprised of longshoremen and financial analysts would have more variability than our data). University research projects are an important application because they train the highly-skilled workforce that is valued by high-tech, high-growth firms (Zolas et al. 2015). In this sense, our work sets the stage for quantifying important forms of intangible (human) capital, which have proven challenging to measure (*24*).

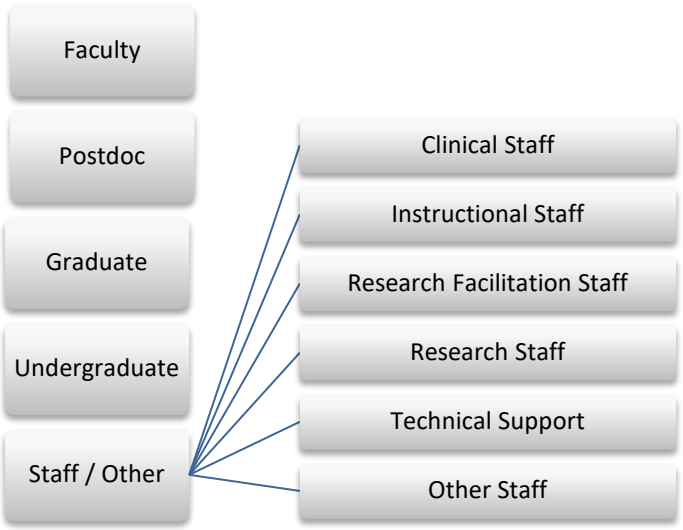
In determining occupational classifications, we drew heavily on the principles enunciated by the federal agencies. We were particularly interested in building a classification system that described the way in which people are used in the production of research. Our classification system benefited from extensive consultation with universities, which identified five core characteristics that distinguish personnel employed on research projects: (i) Permanence in their position (ii) Research Role, (iii) Professorial Track, (iv) Scientific Training and (v) Clinical Association. These core characteristics are similar to ones used in Standard Occupational Classification (SOC) system: classification principle #2 reads "Occupations are classified based on work performed and, in some cases, on the skills, education, and/or training needed to perform the work at a competent level."

Based on this input, we iteratively developed a hierarchical occupation classification system. In the end, we identified a two-level classification system. The first level is based on a person's relationship to the university – faculty, undergraduate, graduate student, postdoc, or staff/other. In the second level, we subdivide staff/other based on function. Figure 1 lays out our classification system and Appendix I provides illustrative job titles for the occupations.

As we discuss in detail in the following sections, we manually assigned an occupation from our classification system to job titles from the eight universities. Then, we used this manually curated data linking job titles to occupations as a training dataset for a supervised machine learning approach that algorithmically assigns occupations to job titles.[5]

Figure 1. Classification System.

[5] Our sample consists of individuals appearing in the employee file between 2012 and 2014. Universities that are missing records in any year between 2012 and 2014 were dropped. Universities that had less than 100 employees in any occupational class were also dropped because the accuracy of classification algorithms may not be reliably calculated. There were eight universities that satisfied these sample restrictions.



IV. Creating a Training Dataset from HR Records

The first step was to manually classify occupations based on job titles, which points to the scale of the problem and hence the value of an automated approach. First, the total number of job titles varied from the low hundreds to low thousands across universities – it is likely that similar variation occurs in firms in other sectors of the economy.

The composition of the research personnel by occupation is shown in Table 1.[6] Also shown in Table 1 is the average number of person-years by occupation for the four largest and four smallest universities (i.e., those universities whose total number of person-year counts is above or below the median). Big universities have, on average, twice as many research personnel paid by research grants and the share of graduate and undergraduate students is somewhat larger for the big universities.

Table 1. Number of employees paid by research grant by occupation.

[6] The occupational Staff and Others were combined into a single category because the distinction between the two classes are somewhat ambiguous and less important. The unit of observation is a person-year. That is, an individual can be counted up to three times, once per calendar year. Because career transitions can happen within a calendar year (e.g., an individual changing his or her occupation from graduate student to postdoctoral researcher over the summer), only individuals who appeared in the employee table under the same job title both before July 1 and after September 30 were included in our sample.

Occupation	All universities Total	Big universities Average	Small universities Average
Faculty	16,000	2,600	1,500
Graduate	17,000	3,100	1,200
Staff / Other	29,000	4,700	2,600
Postdoc	6,900	1,100	650
Undergrad	9,700	2,000	450
Total	79,000	13,000	6,400

Note. The table shows the number of employees paid by research grants at all universities in our data and those with more than and fewer than the median number of person-year pairs. Numbers are rounded for disclosure protection reasons.

It is also worth noting that there is substantial variation in the number of people with each job title, as reflected in the average number of people per job title and the fractions of job titles that contain different numbers of people. We divide universities into two groups – the four with the "coarsest" and the four with the most "detailed" job titles. As shown in Table 2, for the universities that use more detailed job titles, as much as 30% of job titles had only one employee. For the universities that use coarse job titles, the proportion occupied by the job titles with more than 100 employees is nontrivial, and job titles with more than 1000 employees were not uncommon. This has important implications for our work – the handling of some job titles has much more effect on accuracy of the entire occupation classification than others.

Table 2. Variation in the volume and size of job titles across universities.

	All universities	Universities with coarse job titles	Universities with detailed job titles
Total number of job titles (across universities)	3,200	1,100	2,200
Total number of employees (across universities)	79,000	48,000	31,000
Average # employees per title (at each university)	24.4	44.4	14.5
1 employee	25%	16%	30%
2-10 employees	54%	52%	55%
11-100 employees	17%	26%	13%
>100 employees	4%	7%	2%

Note. The table shows the distribution of the number of employees per title at all universities in our data and the 4 universities with the smallest and the 4 universities with the largest numbers of employees per job title.

Even using the relatively straightforward categorization depicted in Figure 1, we identified three separate measurement challenges that will almost surely be manifested in other firms across the economy. Each results in issues that affect the quality of the training data.

First, when different employees with the same job title perform different tasks, the same job title can map to two distinct occupations. For instance, consider employees with the job title of "program coordinators". In some cases, these employees may be managing the business operations of a scientific research program at a university center, and should thus be assigned the occupation "Research Facilitation Staff". In other cases, these employees may be involved in educational or student experiences, and should thus be assigned the occupation "Instructional Staff". In this case, different people with the same job title perform different tasks and should thus be assigned to different occupations. This implies that a full classification must operate at the level of individuals rather than job titles.

Second, some job titles are at the margins of categories. For instance, consider employees with the job title "laboratory supervisors". In many cases, these employees appeared to perform some tasks that would suggest assigning them the occupation "Research Facilitation Staff" and other tasks that would suggest assigning them the occupation of "Research Staff". For instance, some laboratory supervisors serve as an administrator for a university research lab and also conduct research within the lab. Because such employees' work encompasses the responsibilities of two occupations, it can be argued that they fall at the margin of the occupational categories, which points to the value of a task / skill-based classification versus a categorical classification. This measurement challenge is conceptually distinct from the first insofar as a single individual performs

functions that cross categories, rather than two separate people with the same job title performing different functions.[7]

The third measurement challenge is ambiguity: vague titles limited our ability to confidently assign occupations to job titles. “Administrative support”, “coordinator”, and “professional aide” are all examples of unclear job titles. Some employees with these titles work in human resources, undergraduate admissions, or a wide range of offices supporting general university functions, while other employees with those titles may be directly involved in supporting or conducting scientific research. To a large extent, this ambiguity reflects a fundamental noisiness in occupational classifications in their own right.[8] When dealing with ambiguous titles, researchers should be aware that it could influence the learning process of machine-learning algorithms if manually classified occupations were subsequently used for training. For example, a job title “student help” can belong to either a student who provides help or a staff member who helps students. If we assign this title to a student occupation, we implicitly reinforce the association between the word “student” in the job title and the title belonging to a student occupation,

---

[7] Although jobs at the margins of categories are not limited to managerial jobs (and our categories are carefully chosen to minimize such uncertainty), managerial jobs often lie at the margins of categories because they require expertise in different kinds of skillsets. One way to address this issue is to create management occupations as in SOC. For our data, the number of job titles at the margins of categories is relatively small, and therefore, we proceeded without creating managerial occupations, but this will likely pose a substantial problem outside the research industry.

[8] We benefited tremendously from input from member universities who provided extensive input on our classification approach up front; provided a wealth of data; and who have, in many cases, provided extensive feedback on our classification of their employees, especially to address the issues above. One issue that arose in our consultation process is that different universities classify the same workers differently and sometimes in idiosyncratic ways. For example, a few institutions classify all librarians or lecturers as faculty. In implementing this system, we have opted to impose a uniform classification system (e.g. classify librarians and lecturers as staff across campuses) to maximize comparability. This points to the limits of relying on institutions to classify their own data.

potentially increasing the chance of misclassification for job titles such as “student learning center coordinator”. Another example of this type is “fellowship”, which may be intended to mean “fellow”, usually a graduate student, or a staff who handles administrative work involving fellowship. Addressing title ambiguity is conceptually straightforward, but it requires a great degree of cooperation from data-submitting organizations.

It is worth noting that the same person can have multiple relationships to a university. For instance, a student may hold a staff position or a staff can become a student to take advantage of a discount on tuition. In this case, the person would be both a staff member and a student. Such multiple relationships pose a challenge, but also present an opportunity for obtaining unique data on career paths. The ideal handling of such cases depends on the intended use of the data. If one wants to measure the inputs to a production function, then the preferred approach would likely be to assign the person to the staff title (i.e. to the role that they are playing on the sponsored project in question). If the goal is to identify people who have studied at the university, the preferred approach would be to assign them to the appropriate student occupation. Our data tend to favor the first approach because the primary classification is based on the job title.

Another issue that generates a challenge, but also has the potential to enrich the data greatly, is that people’s relationship to a university may change over time. An undergraduate may graduate and enter a graduate program at the same school or take a job as a staff member. A graduate student may take a staff, faculty or postdoc position upon completion of their degree. Obviously, some such pathways are more likely than others. These transitions potentially provide additional leverage on the classification of specific job titles and also provide rich data on career paths.

Incorporating additional external information

We use several different sources of external information, including online job descriptions, publicly available electronic salary databases, university and professional networking websites, and historical administrative earnings and employment data.

Many firms will have HR descriptions that map directly on to job titles. This information could, in principle, provide substantial external information that can be leveraged for occupational classification. In our case, the eight universities had searchable databases for employment and job postings on university HR websites. These typically provided detailed descriptions of specific job titles to confirm the nature of an employee's work. When these descriptions failed to provide the necessary information to correctly classify a position, electronic salary databases for public universities proved to be particularly helpful sources of information on employee names. Using name and job titles enabled us to examine individual profiles on university and professional networking websites, both of which offered detailed explanations of employees' work. Specific information on actual employees rather than just their titles enabled a more careful classification of similarly related positions in some cases.

Placement and earnings are obtained by linking UMETRICS data to data at the U.S. Census Bureau. Given large differences in age and earnings between various occupations in our data, information on and individual's job placement and earnings can provide valuable information about that individual's occupation. Employees in the UMETRICS data are linked to Census data using a Protected Identification Key (PIK), Census's internal anonymized individual identifier.[9] Specifically, we use PIKs to link UMETRICS

[9] In order for a Protected Identification Key (PIK), Census's internal individual identifier, to be assigned, the data on the employing university, the employee last name, first name, and (in some cases) date of birth have been provided to the Census Bureau. The Census Bureau's Person Identification Validation System (PVS) is used to assign an anonymous, unique person identifier to university employees (*26*). UMETRICS employee name, address, and date of birth, when available, are parsed, standardized and geocoded during the input process for the PVS. Next, a probabilistic match is performed between the UMETRICS data and PVS reference files that are based on the Social Security Administration's Numerical Identification File (Numident). When possible, PVS assigns the person PIK. Because PVS is a probabilistic match, it is possible for a UMETRICS employee to receive multiple PIK values. UMETRICS employee data is historic and spans multiple years. Thus, a custom PVS process with many years of associated reference files for each university is used. For detailed information about reference files in PVS or the matching algorithm, see Wagner and Layne 2014)

employees to W2 and LEHD (Longitudinal Employer Household Dynamics) data, from which we obtain earnings and the EIN (employer ID number) of the firm at which they are employed. We are then able to use the PIK-EIN to link UMTRICS employees to the Business Register (BR), the Longitudinal Business Database (LBD), and the Integrated Longitudinal Business Database (iLBD). This enables us to track the job market outcomes of the researchers paid by federal grants and the location, characteristics, and performance of the firms they work for.[10]

V. Measurement and Standardization

The development of clear standardized protocols for interviewers is critical for consistent measurement across individuals. Similarly, good measurement is critically dependent on developing consistent protocols for preprocessing the data so that measures can be standardized across businesses. This is particularly important since each business will have different shorthand to classify job titles. In this section, we will illustrate the challenges of standardizing data collected across multiple organizations with different conventions. We will focus on the abbreviated nature of job titles, but we expect similar

---

Not all universities provide employee date of birth, resulting in higher rates of multiple PIKs than when date of birth is present. A filter is applied to all university employee PIKs in order to select the correct PIK from the multiple values when possible as well as to screen false one-to-one matches. W-2 data used for the filter is limited to records for the years over which the university employee data spans, the EIN(s) associated with the university, and addresses within a 200-mile radius of the university campus address. For each university, PIK values associated with the UMETRICS data are looked up in the W-2 data to create a linkage between UMETRICS data and W-2 data. A match to the W-2 data must occur for that employee to be retained in the sample. For multiple PIK values, only the PIK that appears in the W-2 data is retained for the employee. Filtered data are output to the employee crosswalk data file.

[10] About 20% of the doctoral recipients in our data are not matched to LEHD dataset. This can be for several reasons: (i) the recipient does not have a job in the US – either for family reasons or because they go back to their home country, (ii) they start up a business rather than choose employment or (iii) it is not possible to uniquely match them to a PIK.

challenges will arise in processing texts describing job responsibilities, salary grade, retirement benefits, and other information that may be available.

To automate the classification process, we first need to convert job titles to numeric values because most machine-learning algorithms accept only numerical inputs. For short texts like job titles, the most common way of converting texts to numeric features (equivalent of regressors in regression analysis) is to record the presence/absence of keywords. For example, if we have job titles “research analyst” and “research support”, the array of feature names is [“research”, “analyst”, “support”] and the text-to-feature conversion would return the vector [1, 1, 0] for “research analyst” and [1, 0, 1] for “research support”. These vectors will then be used as inputs for machine-learning algorithms trying to predict occupations.

One problem with this approach is different abbreviations/synonyms in the job titles that represent the same feature. For example, it is clear to humans that “assistant” and “asstnt” both represent “assistant”, but machines treat them as different features. To avoid creating separate features for different abbreviations of the same word, job titles need to be normalized before being converted to numeric vectors.

Because creating a normalization mapping is labor intensive, one may be tempted to use edit distance to determine if a string of letters is an abbreviation of a word. However, generic edit distance fails to address challenges that are specific to abbreviations: for instance, both “busin” and “buses” are formed by deleting three letters from “business” and therefore have the same edit distance; however, the former is more likely to be an abbreviation for “business” than the latter. Developing a set of rules for determining the validity of abbreviation is not a trivial task. Though the disabbreviation algorithm we developed is imperfect, we employed the algorithm for the subsequent analyses to reduce noise in the data (see appendix for details).

VI. Machine Learning

We first conducted a preliminary analysis comparing the performance of Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Random Forests, and Extra Trees and found that

random forests best fit our purpose. Random forests construct a multitude of decision trees from training data and assign the mode class (occupation) to a new observation (job title). Random forests mitigate decision trees' tendency to overfit by adding randomness in both the sample observations used (bagging) and the set of features considered at each node split.

Figure 2 shows part of a decision tree that classifies employees into the main five classes from Figure 1 (faculty, postgraduate students, graduate students, undergraduate students, and staff / other) based on their job titles. Each box contains (i) branching rule; (ii) Gini impurity; (iii) number of observations contained in the node; (iv) composition of observations; and (v) majority class.

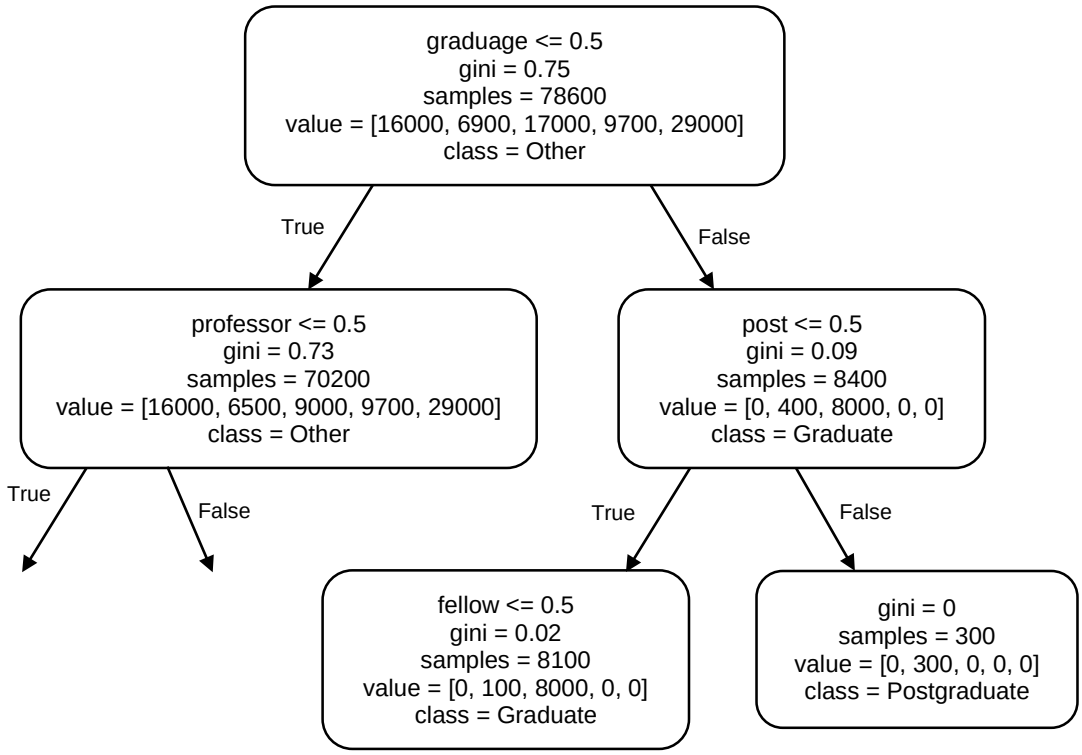


Figure 2. Example of a decision tree. Each node has a keyword indicated at the top of the box. All observations that have the keyword in their job titles follow the right branch while observations without the keyword follow the left branch. When an observation reaches a terminal node like the one on the right bottom, the class of the node becomes the predicted class for the observation.

Branching rules specify the feature name and the cutoff value. For example, at the top node, job titles that contain the word “graduate” less than or equal to 0.5 times follow the left branch, while those that contain the word “graduate” more than 0.5 times follow the right branch. Because feature values are integers, it is equivalent to the following: job titles without the word “graduate” follow the left branch and those with the word “graduate” follow the right branch. This tree is intuitive. If “graduate” is not present, it next tests for “professor” and if “graduate” is present, it tests for “post” (as in postgraduate). Note that the node at the bottom right does not have the branching rule because it is a terminal node.

“Samples” represents the number of observations in each node. For example, the top node contains 78,600 observations. Of the 78,600 observations, 70,200 observations follow the left branch and 8,400 observations follow the right branch. “Value” shows the composition of the observations: the top node consists of 16,000 faculty, 6,900 postgraduate students, 17,000 graduate students, 9,700 undergraduate students, and 29,000 staff / other. For this node, the majority is “staff / other”, and therefore, the “class” for the node is “staff / other”.

Finally, “gini” reports Gini impurity. Notice that Gini impurity decreases as one goes down the tree and attains 0 when a node consists of one class (bottom right node). It is calculated as follows:

$$G = \sum_{c=1}^{5} p_c(1 - p_c),$$

where p_c is the proportion of class c observations at the node. For example, for the first node, the proportion of faculty, postgraduate, graduate, undergraduate, and staff / other are 0.204, 0.088, 0.216, 0.123, and 0.369, respectively, and the corresponding Gini impurity is

$$G = 0.204(1 - 0.204) + 0.088(1 - 0.088) + 0.216(1 - 0.216) + 0.123(1 - 0.123) + 0.369(1 - 0.369) = 0.753.$$

In our analysis, we will use the predictive accuracy as a measure of performance. Formally, the accuracy is define as

$$Accuracy = \frac{\#(predicted\ class == true\ class)}{\#\ total\ observations}.$$

Throughout our analysis we always hold out data from one university, one at a time, for testing and use data from the remaining seven universities for training. Thus, for a given set of tuning parameters (discussed below), we grow eight separate random forests, each using data from one university for testing the accuracy of the forest and using data from the other seven universities for training the random forest. Thus, for each job title, we have eight occupation predictions, which we aggregate using the mode, into a single occupation for that job title.

Random forests have three main tuning parameters: 1) the total number of features supplied to the random forest, 2) the number of features to be considered at each node of the tree, and 3) the number of trees grown in the forest (i.e. the number of samples randomly selected to build a decision tree). The tradeoff in including more features overall is between having more features to improve prediction and overfitting because of idiosyncratic relationships that may be present in the data. We filter out noise in sample by pre-selecting the features to avoid overfitting idiosyncratic relationships that may be present in the sample too much. The total number of features used in the random forest controls the amount of noise to avoid overfitting. The number of features that the random forest can choose between at each stage controls the variability of the trees: the smaller the set of features to be considered, the more variable the trees become because there is more randomness in the selection of the feature. In the extreme case where only one feature is considered at each split, the selection of feature is totally random (i.e. whatever feature is selected becomes the one used).

To determine the total number of features supplied to the random forest, we fit a decision tree, for each training set, using all 1-grams and 2-grams that appeared in the job titles.

Then, the feature importance score was calculated[11], and the features with the highest importance scores were selected, varying the score cutoff. The total number of features[12] that were fed into the model varied depending on which university was reserved for testing, but was roughly 50, 100, 200, 500, and 7000, where 7000 is the total number of 1-grams and 2-grams appearing in the job titles in the training set and 500 is the number of features that had a strictly positive importance score. We also varied the the number of features considered at each split (default is the square root of the total number of features supplied to the random forest). Finally, we varied the number of trees grown in the forest, in increments of 100, between 100 and 1,000.

In determining the optimal parameter setting, we considered both unweighted and weighted accuracy. The unweighted accuracy was computed treating each job title as one observation; no matter how many employees have that job title, the title receives a weight of 1. The weighted accuracy was computed treating each individual as one observation; equivalently, job titles were assigned a weight equal to the number of employees that have that job title. The most important tuning parameter for determining classification accuracy was the total number of features provided to the random forest (which is implicitly determined by the importance score cutoff). The fraction of features to be considered at each node and the number of trees grown had a minimal effect on the accuracy. Based on the overall weighted and unweighted accuracy, the optimal parameter setting limits the number of features supplied to the random forest to about 200 and uses the default setting of the square root of the total number of features to be considered at each node.

---

[11] We used the DecitionTreeClassifier in the SciKit Learn package.

[12] There were total of 7000 features. It is common to pre-select important features. In one specification (cutoff = 0), all the features were fed to the random forest. In the most selective case, only 50 features were fed to the random forest.

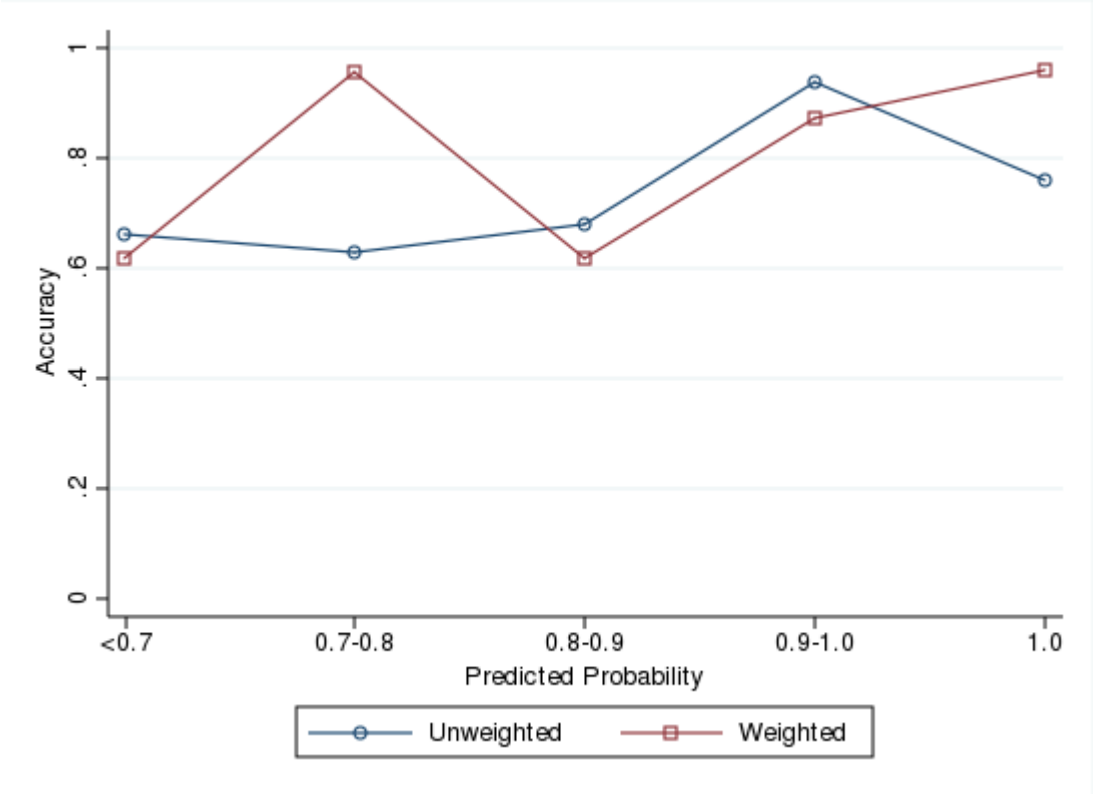


Figure 3: Classification Accuracy relative to Predicted Probability. The figure shows the probability that an occupation was correctly coded as a function of the probability that the algorithm predicts it was correctly coded. The unweighted series treats job titles as the unit of observation. The weighted series treats individuals as the unit of observation.

Figure 3 shows the accuracy (proportion of correct prediction) for each level of predicted probability (the probability share of the predicted occupation indicated by the posterior distribution returned by the algorithm). The overall accuracy varied from 60% to nearly 100% regardless of whether the data are weighted by number of job titles or individuals[13]. Although random forests can potentially increase the efficiency of occupational classification, an average accuracy of about 80% may not be high enough to justify a total replacement of manual classification by automated machine-learning algorithms. These results reinforce our belief that the predicted probability the number of

[13] Unweighted accuracy is the proportion of job titles whose predicted class matched the true class. For weighted accuracy, the number of employees for the job title is used as a weight.

individuals that hold a job title should be used to jointly identify job titles for manual review.

We see two (potentially complementary) roles for machine learning in occupation coding and other similar bucketing tasks. One approach is to use an algorithmic approach to classify uncommon job titles. Such cases are (by construction) plentiful and have a relatively small effect on the overall accuracy of the classification. The second role is to accept only predictions with concentrated probability mass at one class. In other words, we will adopt the prediction only when the random forest classifier is "confident". Obviously, these two approaches could be combined – defining "isoquants" over the size and accuracy to trigger manual review. In this approach only relatively large, uncertain job titles would be reviewed manually.

Robustness

We have explored a wide range of modifications of our basic approach to try to obtain performance improvements. We outline the analyses we performed here and their main results. Appendix II provides details on both the analyses and their results.

For the eight universities used in the above analyses, the number of employees ranged roughly from 5,000 to 20,000. When we train the random forest classifier on seven universities, it is possible that the shape of a tree is heavily influenced by a few universities in the training set with a large number of employees. To investigate this possibility, the training set was modified so that universities in the training set have roughly equal numbers of employees. The modifications have been made in two ways: inflating and deflating. As Section A in Appendix II shows, there was no significant change in the accuracy with these modifications.

The number of employees per job title ranged from 1 to nearly 10,000 for the eight universities, with the average being 24.4 employees per title. Concerned that the titles in the training set are "too noisy", we investigated the effect of dropping thin titles (varying the threshold at which a title is flagged as "thin" from 5 to 50 employees) from the

training set. We recorded the average predictive accuracy for titles with different numbers of employees. Again, there was no significant change in the accuracy with these modifications. These results are discussed in Section B of Appendix II.
We observed that some titles that could be easily classified manually like "Graduate Assistant" are not always correctly classified by our random forest. This appears to be caused by the existence of "extraneous" information in some job titles. In Section C, we applied partially unsupervised learning. In particular, titles that (after applying the job cleaning algorithm outlined in the appendix) contain the words "faculty", "professor", "postgraduate", "graduate" or "undergraduate" were classified first and then the random forest classifier was applied to the remaining titles (both the training set and the test set consist of titles that do not contain any of the words listed above). The effect of this partially unsupervised learning on the predictive accuracy is small, with our classification for some universities improving and others degrading.

Census Bureau links permitted us to examine whether having information on individuals' age and earnings increased the quality of prediction. These variables would appear to be valuable predictors, especially in this context because of the large differences in ages and earning across occupations. As shown in Table A4, there is some gain, but it is not extraordinarily high across the board. The largest gains, by far, are for undergraduates when occupations are weighted by the number of people in them.[14]

As indicated in the previous sections, people can hold multiple titles at a point in time (or in close succession) and can transition between titles. As some transitions are more common than others (i.e., transitions from undergraduate to graduate and/or from graduate to postgraduate and/or from postgraduate to faculty are more common than the reverse transitions), it is possible to use transitions between titles and concurrent titles (more precisely, occupational classes that are associated with these titles) as predictors in

[14] Because there are considerably more staff than undergraduates overall, there is a tendency for the random forest to misclassify undergraduates as staff.

the random forest classifier to improve predictive accuracy. Transitional and concurrent titles can also be used to identify unlikely transitions in the "ground-truth" data, providing an opportunity for a revision. Beyond improving the accuracy of the data, exploring concurrent positions and transitions can add to the richness of our data by providing information on career paths. Section D provides details of this analysis.

Using concurrent job titles and the transitions between job titles involves some form of iterative procedure. Appendix II details a number of issues related to using transitions and concurrent titles. As a first step toward incorporating transitional and concurrent classes in the random forest classifiers, we included the manually classified transitional and concurrent classes in our training data in the model rather than predicted occupations. The resulting predictive accuracy is expected to provide an upper bound for the accuracy obtained from the iterated procedure described above. Overall, the use of concurrent titles and transitions across titles has little effect on overall accuracy. In our analysis, no university exhibited a clear pattern on the effect of including transitional/concurrent class as predictors.

Limitation of Machine-Learning Algorithms

Laying aside the issue of developing a classification system, we have discussed four challenges to manual classification. Beyond these issues associated with manually classifying occupations, comparing the predictions made by the random forest and the true class pointed to two possible causes of misclassification. One is unavoidable misclassification, which results from variation in the training data. The other is avoidable misclassification, which results from the inherent limitations of the random forest classifier.

The first type of misclassification is unavoidable because it arises from the limits to manual classification already discussed. Listed below are examples of job titles that have multiple classifications over universities.

- Director, dean: faculty or staff
- Instructor, lecturer: faculty, graduate, or staff
- Grader, tutor: graduate or undergraduate

- Scientist: faculty or staff / other
- Research assistant: graduate, undergraduate, or staff / other
- Research associate: postgraduate or staff / other
- Fellow: postgraduate, graduate, or staff / other
- Intern: postgraduate, undergraduate, or staff / other

This type of inaccuracy cannot be overcome by any classifier: resolution of misclassification requires familiarity with job titling convention at each university. It should also be noted that modifiers can change the classification of a job title within a university. For example, “director” and “associate director” may not belong to the same category within a university.

The second type of misclassification is avoidable. Avoidable misclassifications are due to the limitations of the random forest classifier. Below are examples of misclassified job titles along with the prediction made by the random forest, followed by true class in parentheses.

- Undergraduate fellow → graduate (undergraduate)
- Temporary visiting faculty → staff / other (faculty)
- Teaching assistant → staff / other (graduate)
- Summer term ra (w/o tuit ben) → staff / other (graduate)
- GR AST ½ → staff / other (graduate)

The first two examples illustrate the tendency of the random forest classifier to rely too much on certain words. The word “fellow” is strongly associated with graduate student. Thus, if “fellow” is selected as a branching rule before “undergraduate”, the job title “undergraduate fellow” will be buried in a node that is predominantly graduate students. Similarly, the word “temporary” is often associated with a staff member and almost never used for faculty. The partially supervised machine learning algorithm described in the previous section is intended to address these issues.

The third example illustrates failure to utilize very informative words or phrases. The presence of the phrase “teaching assistant” in a job title is a good indicator of the employee being a graduate student. However, the absence of the phrase “teaching

assistant" in the job title is not a good indicator of the employee not being a graduate student (i.e., there are many graduate students who are not teaching assistants). Thus, when the phrase "teaching assistant" is used for branching, the resulting decrease in impurity of the succeeding node is negligible. Since the random forest classifier selects the feature that minimizes the weighted average of impurities at succeeding nodes, the phrase "teaching assistant" is unlikely to be selected.

The last two examples illustrate inability of the random forest classifier to use outside knowledge. A human classifier can infer "w/o tuit ben" means "without tuition benefit" and conclude that the job title is associated with a student. Similarly, "1/2" suggest that the person has a half-time appointment, and therefore likely be a student. Thus, one may infer that "gr ast" means "graduate assistant". As seen in the previous example, these phrases are extremely informative; however, because of their rare occurrence and applicability to only a small fraction of employees, these pieces of information tend to be overlooked by the random forest classifier.

In theory, misclassifications described above might be reduced by providing more training data, adjusting parameters, appealing to other machine-learning algorithms, or reverting to manual classification.

VII. Conclusions

This paper used a rich dataset – to our knowledge, the first dataset with detailed job titles drawn from HR systems from multiple organizations, combined with job descriptions and information about the characteristics of workers – to examine the potential to use machine-learning techniques for occupational classification. We followed the same conceptual framework as that applied by survey methodologists: to define each occupation, to translate concepts to standardized protocols and to build an approach that would infer occupations from the information at hand. Even though the data were drawn from very similar organizations, with very similar production functions, we found that machine-learning approaches were not substantially better than manual classifications.

However, we do see the analysis as showing real promise for identifying occupations from job titles combined with a machine-learning approach. The most promising use of the machine learning is that it is an inexpensive way of assigning occupations for job titles that have relatively few people in them and/or for which the algorithm imputes a high degree of accuracy. Because many job titles have only a few people in them, this approach could yield substantial cost savings (almost 80% of job titles have 10 or fewer people). At the same time, an entirely algorithmic approach would be unwarranted in our case.

We also believe that a deeper text analysis of the job descriptions associated with job titles might prove to be a promising approach. Job descriptions typically include information about necessary experience, skills and education which are not only of interest in their own right but could be very useful features for classification purposes.

We note that the focus on universities as a subject of analysis has weaknesses and strengths. Major research universities are very large and complicated institutions. There may be other industries in which it might be easier to apply machine learning to job titles. At the same time, the institutions in our sample all come from one narrow sector of the economy are relatively homogeneous and the data are based on a very specific set of activities (research). We speculate that any classification system for the broader economy would have to be specific to an individual sector or set of sectors.

VIII. References

1. W. Mellow, H. Sider, Accuracy of response in labor market surveys: Evidence and implications. *J. Labor Econ.*, 331–344 (1983).
2. E. Groshen, Innovating for the Future (2017), (available at https://blogs.bls.gov/blog/tag/bls-products-and-services/).
3. Workforce Information Council Administrative Wage Record Enhancement Study Group, "Enhancing Unemployment Insurance Wage Records Potential Benefits, Barriers, and Opportunities" (Washington DC., 2015), (available at https://www.bls.gov/advisory/bloc/enhancing-unemployment-insurance-wage-records.pdf).
4. N. Zolas *et al.*, Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. *Science (80-.).* **350** (2015), doi:10.1126/science.aac5949.
5. D. Autor, F. Levy, R. MURNANE, The Skill Content of Recent Technological Change: An Empirical Exploration. *Q. J. Econ.* **118**, 4 (2003).
6. K. A. Weeden, D. B. Grusky, The Case for a New Class Map 1. *Am. J. Sociol.* **111**, 141–212 (2005).
7. O. Alonso-Villar, C. Del Rio, C. Gradín, The extent of occupational segregation in the United States: Differences by race, ethnicity, and gender. *Ind. relations a J. Econ. Soc.* **51**, 179–212 (2012).
8. D. Acemoglu, U. Akcigit, N. Bloom, W. R. Kerr, "Innovation, reallocation and growth" (National Bureau of Economic Research, 2013).
9. N. G. Peterson *et al.*, Understanding work using the Occupational Information Network (O* NET): Implications for practice and research. *Pers. Psychol.* **54**, 451–492 (2001).
10. C. Levine, L. Salmon, D. H. Weinberg, Revising the standard occupational classification system. *Mon. Lab. Rev.* **122**, 36 (1999).
11. A. Emmel, T. Cosca, The 2010 Standard Occupational Classification (SOC): A Classification System Gets an Update. *Occup. Outlook Q.* **54**, 13–19 (2010).
12. T. Cosca, A. Emmel, Revising the standard occupational classification system for 2010. *Mon. Labor Rev.* **133**, 32–41 (2010).
13. N. A. Mathiowetz, Errors in reports of occupation. *Public Opin. Q.*, 352–355 (1992).
14. J. Bound, C. Brown, N. Mathiowetz, in *Handbook of Econometrics* (2001; http://www.sciencedirect.com/science/article/pii/S1573441201050127), vol. 5, pp. 3705–3843.
15. K. G. Abraham, J. R. Spletzer, New evidence on the returns to job skills. *Am. Econ. Rev.* **99**, 52–57 (2009).
16. J. D. Fisher, C. Houseworth, Occupation Inflation in the Current Population Survey. *US Census Bur. Cent. Econ. Stud. Pap. No. CES-WP-12-26* (2012).

17. Workforce Information Council Administrative Wage Record Enhancement Study Group, “Enhancing Unemployment Insurance Wage Records Potential Benefits, Barriers, and Opportunities” (2014).

18. Texas Workforce Commission, “Report to the Sunset Advisory Commission Study on the Collection of Occupational Data” (Austin, Texas, 2016).

19. A. Bethmann, M. Schierholz, K. Wenzig, M. Zielonka, Automatic Coding of Occupations (2014).

20. K. G. Tijdens, Reviewing the measurement and comparison of occupations across Europe. *AIAS Work. Pap.* (2014).

21. 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).

22. B. A. Weinberg *et al.*, Science Funding and Short-Term Economic Activity. *Science (80-.).* **344**, 41–43 (2014).

23. N. Zolas *et al.*, Wrapping it up in a person: Examining employment and earnings outcomes for PhD recipients. *Science (80-.).* **360**, 1367–1371 (2015).

24. C. Corrado, J. Haskel, C. Jona-Lasinio, Knowledge spillovers, ICT and productivity growth. *Oxf. Bull. Econ. Stat.* (2017).

25. E. A. Fleishman, M. D. Mumford, Evaluating classifications of job behavior: A construct validation of the ability requirement scales. *Pers. Psychol.* **44**, 523–575 (1991).

26. D. Wagner, M. Layne, “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications’ (CARRA) Record Linkage Software” (2014), (available at https://www.census.gov/srd/carra/CARRA_PVS_Record_Linkage.pdf).

Appendix I

1. Detailed Description of Occupations

This section lays out the occupation categories that we use; their conceptual definition; and some illustrative job titles. The aggregate occupations are listed first. Staff are subdivided into additional categories, which are laid out below.

1. Faculty

All advanced academic employees who are directly involved in scientific research and/or scientific instruction. These included: Deans, Provosts, Tenure/tenure track, Clinical, Research, Visiting Professors, Academic specialists, Center directors

2. Post Graduate Research

All individuals holding terminal degrees (PhD, MD) who are in temporary training status. These included: Postdoctoral, Medical residents/interns/fellows, Clinical fellowships, Research Associate (depends on the university)

3. Graduate Student

Students earning advanced degrees: Graduate students (part time, full time), Medical/dental/nursing/students, Research Assistant

4. Undergraduate

Students earning baccalaureate/other degrees including fulltime, part time, summer research assistants, work study; includes high school students who would likely be acting in a similar capacity. These included: Undergraduate students, High school students, Intern/student worker, Nursing students in BA programs

5. Staff / Other (Not Elsewhere Classified)

Positions that support general university functions such as undergraduate education and student activities. Employees whose titles cannot be attributed to the scientific research enterprise. These included at the aggregate level: **Staff** Instructional, Research, Research

Facilitation, Technician, Clinical, Other Staff. The disaggregated staff categories include the following

5.1 Clinical Staff: All non-faculty health care professionals, Nurses (non-faculty), Dieticians (non-faculty), Nutritionists, Social workers, Physical therapists, Clinical psychologists, Dental hygienists, Genetics counselor

5.2 Instructional Academic Specialists: Lecturers, Instructors, Adjunct Professors

5.3 Research Facilitation: Non-faculty, high level administrators – asst. dean/asst. provost, associate or assistant center director, Operations managers/managing directors, Administrative/clerical staff – any kind, Finance staff, Regulatory staff, Clinical or clinical research support staff, Laboratory aide, Data collection/interviewer, Media jobs: Graphics/writer/editor/communications, Grants management & administration, Individuals who serve as managers/ coordinators/facilitators for laboratory studies/clinical trials/large facilities/research programs; They direct and influence scientific research activity from the level of the laboratory up to the level of the university/research center, Research dean/provost/administrator, Facility director/administrator, Clinical research administrator, Study coordinators, IACUC coordinators, Clinical trials/research coordinator, Project/Program manager/coordinator, Lab coordinator (not lab manager), Facility/repository manager/coordinator.

5.4 Research Staff: Work likely focuses on scientific aspects of research All advanced degree qualified, non-faculty scientists and engineers; Research specialist/engineer: Work likely focuses on advanced research analysis; Research professional/specialist; Statistician, bioinformaticist; Research associate (depends on the university); Skilled and specialized employees who have been specifically trained in some area of science & technology; Science Technicians: All technical staff including animal technicians, machinists, mechanics (the category usually includes some reference to a research facility along with the title ‘technician’); Lab manager ; Medical or clinical technician; Research data technician; Regulatory officer (environmental, chemical safety, industrial hygienist); Technical engineer

5.5 Technician: Administrative and technical employees who are not specifically employed for scientific research purposes but perform job tasks that support the research enterprise; Information technology managers & staff; Software engineer; Data entry/data analyst; Network and systems support

5.6 Staff Other All other research staff that do not clearly fall into another category

2. Normalization

We developed a rule-based job title cleaning algorithm. In particular, we created a mapping from abbreviation to normalized word. For example, “grad” is mapped to “graduate” and “mngr” is mapped to “manager”. The list of abbreviations and possible normalized words were obtained from job titles from eight universities in the UMETRICS dataset, and mappings were created manually.

Abbreviations with multiple possible normalizations were noted (e.g., “res” can be an abbreviation for “research” or “respiratory”; “ast” can be an abbreviation for “assistant” or “astronomy”). Then context-specific normalization (i.e., normalization of phrases) was attempted. For example, both “res” and “ast” are ambiguous abbreviations; however, when they are combined, one can infer “res ast” is an abbreviation for “research assistant”. Normalizing rules for phrases were manually generalized using regular expressions.

When an abbreviation could represent either a person or a field (or an object) that are closely related, we chose the field in general. For example, “scien”, “enginee”, and “crimnl” were normalized to science, engineering, and criminology, instead of scientist, engineer, and criminal, respectively. The reason is that it seems more harmful to label non-engineers in engineering departments an “engineer” than to label an engineer “engineering”. When an abbreviation is strongly associated with an occupation, however, we normalized it to represent a person. For example, “lect” and “consul” were normalized to lecturer and consultant instead of lecture and consulting, respectively. These are somewhat ad-hoc rules, but these abbreviations are small in number, and we expect they have a negligible effect on the performance of machine learning algorithms.

When creating the normalization mapping, we preserved common acronyms such as “CSE” for Computer Science and Engineering and “MRI” for magnetic resonance imaging. We expect normalizing these terms has a minimal impact on the predictive accuracy because they identify the fields employees work in, but contain little information on tasks they perform.

At the same time the mapping was created, omissions of spaces were noted and a decomposition mapping was created. For example, we encountered job titles such as “rsrchanalyst”, which was added to the decomposition mapping along with the correction “rsrch analyst”. Common stems in compounds, such as bio in biochemistry and neuro in neurosurgery, were not decomposed and compounds were treated like words.

Finally, on the normalization list, we had some abbreviations that are only 2-letters long. For example, we left “IT” as it is assuming that it represents Information Technology. However, these could be an abbreviation of some other words or phrases. In our data, we did not find any instances where there was a more suitable normalization, but researchers should be aware that too much guessing when standardizing could introduce more noise than it eliminates.

Aside from working out the details, the major problem with the above described normalization algorithm is that the mapping is not comprehensive. For example, “research” may be mapped from “resear”, “rsrch”, and “resch”, but if there is no mapping from “resech” to “research”, “resech” will remain abbreviated. By comparing manually normalized job titles and normalization returned by the algorithm, we identified normalizations that were not captured by the normalization mapping, and iteratively revised our normalization mapping. We also wrote regular expressions to normalize words that frequently appear in our data such as “research”, “postdoctoral”, and “administrator”.

3. Coding decisions

There are also some methodological issues of interest. First, we designed our classification to increase certainty: grouping workers whose job was sufficiently similar that it would be hard to separate them based on job titles (and for whom the value of distinguishing occupations has the least value). Second, we employed a two-level system, where the first-level occupation can frequently be assigned with a high degree of certainty and much of the uncertainty appears at the second level. Third, we assigned up to two occupations to each job title to allow researchers to probe the sensitivity of results. Fourth, we rate job titles based on the degree of certainty that they were correctly classified on a scale of 1-5. Our coding system was:

(5) The job title serves as an immediate identifier into this classification category or, through research, it is almost certain that it belongs in this category: e.g. PostDoctoral Fellow; Computer Technician

(4) The job title probably belongs in the category indicated, as supplemented by research on university website.

(3) The job title belongs in the category (either aggregate or disaggregate) with moderate certainty (either very indicative job title or research result, but not both)

(2) The job title is vague and/or ambiguous but there is some indication that the position belongs in this category.

(1) The job title may belong in this category, but there is little certainty, and the classification cannot be verified with research.

After manual classification, universities were given the opportunity to review and comment on the classification, with their attention drawn to the largest and most ambiguous titles.

Appendix II

A. Different Numbers of Employees

For the eight universities used in the above analyses, the number of employees ranged roughly from 5,000 to 20,000. When we train the random forest classifier on seven

universities, it is possible that the shape of a tree is heavily influenced by a few universities in the training set with a large number of employees. To investigate this possibility, the training set was modified so that universities in the training set have roughly equal numbers of employees. The modifications have been made in two ways: inflating and deflating.

(1) Inflating

Let

$N_{u,t}$ = number of employees at university u for job title t, and

N_u = number of employees at university u.

Then the modified number of employees is

$$\widetilde{N}_{u,t} = N_{u,t} \times \frac{\max_v\{N_v\}}{N_u},$$

rounded to the nearest integer. For example, if university X has a total of 16,000 employees and if the largest university in the training set has a total of 20,000 employees, the number of employees for each title at university X is multiplied by 1.25 and rounded to the nearest integer. If a title has 3 employees, the inflated number of employees is $1.25 \times 3 = 3.75$, so it will be rounded to 4.

(2) Deflating

Instead of scaling up the number of employees to the level of the largest university in the training set, deflating scales down the number of employees to the level of the smallest university:

$$\widetilde{N}_{u,t} = N_{u,t} \times \frac{\min_v\{N_v\}}{N_u}.$$

For example, if university X has a total of 20,000 employees and if the smallest university in the training set has a total of 5,000 employees, the number of employees for each title at university X is multiplied by 0.25 and rounded to the nearest integer. If a title has 10 employees, the deflated number of employees is $10 \times 0.25 = 2.5$, so it will be rounded to 3. If a title has 1 employee, the deflated number of employees is

$1 \times 0.25 = 0.25$, so it will be rounded to 0. In other words, the title will be dropped from the training set.

Result

As evident in Table A1, inflating and deflating the number of employees in the training set has no effect on the unweighted accuracy. There is a little improvement in the weighted accuracy for big universities when the number of employees in the training set is deflated. One possible explanation is that deflating reduces the noise in the training data because uncommon job titles are dropped from the training set due to rounding if the deflated number of employees is less than 0.5.

Table A1. Accuracy when total weight is balanced across universities

Size of university	Weight	Benchmark	Inflating	Deflating
All universities	unweighted	0.83	0.83	0.82
Big universities	unweighted	0.87	0.86	0.86
Small universities	unweighted	0.80	0.80	0.79
All universities	weighted	0.84	0.82	0.85
Big universities	weighted	0.83	0.82	0.86
Small universities	weighted	0.84	0.82	0.82

B. Discarding Thin Titles

The number of employees per job title ranged from 1 to nearly 10,000 for the eight universities, with the average being 24.4 employees per title. Concerned that the sparsely populated titles in the training set are particularly "noisy", we investigated the effect of dropping thin titles from the training set. The question we try to answer is "Do thin titles negatively affect the learning and consequently degrade the performance of predicting for heavily populated titles?"

For each university, we used the remaining seven universities for training and discarded the titles with less than a certain number of employees in them from the training set. We used the threshold of 5, 25, and 50 employees per title. Then we recorded the average predictive accuracy for titles grouped by the number of employees per title: 1–4 employees, 5–24 employees, 25–49 employees, 50–99 employees, 100–499 employees,

500–999 employees, and 1000+ employees. The idea is that dropping titles that have fewer than a certain number of employees from the training set may have different effect on the prediction accuracy for thin titles and that for heavily populated titles.

Result

The resulting prediction accuracies are shown in Table A2. Cutoff = 0 corresponds to the benchmark, where all job titles are included in the training set. As the cutoff value increases, more and more job titles are excluded from the training data. There is a small decrease in accuracy caused by discarding uncommon titles for job titles that are relatively small. In contrast, the accuracy improves as the cutoff increases for job titles with 1000 or more employees. This is expected because highly populated job titles tend to have simple, straightforward descriptions and therefore do not benefit from infrequently used features brought to the training set by uncommon job titles. Indeed, excluding uncommon job titles from the training set makes the training set less noisy, allowing the random forest classifier to construct better decision trees with a high predictive accuracy.

Table A2. Accuracy when thin titles are discarded from the training set

Size of Title	Weight	Cutoff = 0	Cutoff = 5	Cutoff = 25	Cutoff = 50
<5	unweighted	0.81	0.81	0.82	0.80
5-24	unweighted	0.84	0.84	0.84	0.82
25-49	unweighted	0.91	0.91	0.91	0.89
50-99	unweighted	0.88	0.87	0.86	0.84
100-499	unweighted	0.82	0.82	0.85	0.82
500-999	unweighted	0.88	0.88	0.88	0.88
>=1000	unweighted	0.75	0.83	0.92	0.92
<5	weighted	0.82	0.82	0.83	0.81
5-24	weighted	0.83	0.83	0.83	0.81
25-49	weighted	0.92	0.91	0.92	0.89
50-99	weighted	0.88	0.87	0.86	0.84
100-499	weighted	0.80	0.79	0.82	0.80
500-999	weighted	0.90	0.90	0.90	0.90
>=1000	weighted	0.79	0.85	0.93	0.93

C. Partially Unsupervised Learning

After observing that titles like “Graduate Assistant” are not always correctly classified, we applied partially unsupervised learning. This appears to be because of “extraneous” information in some job titles. In particular, titles that contain the word (after applying the job cleaning algorithm) “faculty”, “professor”, “postgraduate”, “postdoctoral”, “graduate” or “undergraduate” were classified first and then the random forest was applied to the remaining titles (both the training set and the test set consist of titles that do not contain any of the words listed above).

The resulting accuracies are shown in Table A3. There is no difference in the unweighted accuracy between the supervised learning benchmark and partially unsupervised learning. Contrary to our expectation, the weighted accuracy deteriorated for the universities with granular job titles while it improved for the universities with coarse job titles. This suggests a possibility of overfitting; in the absence of very important features such as “faculty” and “undergraduate”, less important features appear to be more important than they actually are. One possible solution is to recalibrate the parameters to filter out marginally informative features.

Table A3. Accuracy using Partially Supervised Learning

University	weight	benchmark	partially unsupervised
all universities	unweighted	0.83	0.83
universities with coarse job titles	unweighted	0.90	0.91
universities with granular job titles	unweighted	0.79	0.79
all universities	weighted	0.83	0.84
universities with coarse job titles	weighted	0.82	0.86
universities with granular job titles	weighted	0.84	0.81

D. Using Age and Wage Data

Census Bureau links permitted us to examine whether or not having information on individuals’ age and earnings increased the quality of prediction. These variables would appear to be valuable predictors, especially in this context because of the large differences in ages and earning across occupations. As shown in Table A4, there is some

gain, but it is not extraordinarily high across the board. The largest gains, by far, are for undergraduates when occupations are weighted by the number of people in them. The benchmark analysis shows the predictive accuracy for all individuals whose true occupation falls in the occupation indicated in the row heading. The column headed "age and wage" shows the predictive accuracy for individuals for whom we have age and wage information (i.e., subset of the benchmark population). For this subset of population, the "age" column shows the accuracy when age is used along with job title for prediction; the "wage" column shows the accuracy when wage is used along with job title for prediction; and the "age and wage" column shows the accuracy when both age and wage are used along with job titles for prediction.

One interesting observation is that the predictive accuracy increases drastically for graduate students with age and wage information. This may be because common jobs titles like teaching assistant are associated with more standardized hiring procedures that increase the chance of students' information being stored in a more organized way on the university system. This effect, however, disappears when the job titles are not weighted by the number of employees associated with that job title (the bottom half of the table). This could be due to idiosyncratic job titles and associated non-standardized hiring process.

The table also shows that correctly classifying undergraduate students is particularly difficult even with the information on age and wage. This is probably due to heterogeneity within the undergraduate researcher body: there are traditional students straight out of high school as well as adult students whose study is financed by the company they work for.

Table A4. Accuracy using age and wage data

Fraction of individuals whose predicted class matches the true class by true class					
Actual Occupation	Benchmark	Sample with Age & Wage Data	Using Age Data	Using Wage Data	Using Age and Wage Data
Faculty	0.81	0.87	0.88	0.88	0.89
Graduate	0.09	0.73	0.73	0.73	0.73
Staff / Other	0.97	0.96	0.96	0.94	0.94

Postdoc	0.87	0.57	0.63	0.70	0.63
Undergrad	0.23	0.17	0.36	0.39	0.37
Overall	0.65	0.82	0.84	0.84	0.84
Fraction of job titles whose mode of predicted classes matches the true class by true class					
Faculty	0.81	0.90	0.90	0.92	0.92
Graduate	0.09	0.13	0.12	0.12	0.12
Staff / Other	0.97	0.95	0.97	0.95	0.96
Postdoc	0.87	0.85	0.87	0.86	0.86
Undergrad	0.23	0.10	0.08	0.10	0.09
Overall	0.65	0.78	0.79	0.79	0.79

E. Transitional and Concurrent Titles

As indicated, people can hold multiple titles at a point in time (or in close succession) and can transition between titles. As some transitions are more common than others (i.e., a transition from undergraduate to graduate to postgraduate to faculty is more common than the reverse set of transitions), it is possible to use transitions between titles and concurrent titles (more precisely, occupational classes that are associated with these titles) as predictors in the random forest classifier to improve predictive accuracy. Transitional and concurrent titles can also be used to identify unlikely transitions in the "ground-truth" data, providing an opportunity for a revision. Beyond improving the accuracy of the data, exploring concurrent positions and transitions can add to the richness of our data by providing information on career paths.

Using concurrent job titles and the transitions between job titles requires some form of iterative procedure. Obviously, the complete mapping between the set of job titles to itself is too high dimensional to be of any practical use. Thus, we use the following approach. In the first iteration, we predict occupational class using only job titles as predictors. In the second iteration, the predicted classes of the transitional/concurrent titles from the first iteration are used as predictors, along with the job titles. In principle, this process could be iterated until the prediction converges according to some criterion. The example below, where an individual held three job titles in sequence, illustrates our approach.

Table A6. Illustration of iterative procedure using transitions

First Iteration:

Job Title	**Preceding Class**	**Concurrent Class**	**Succeeding Class**	**Prediction**
Student Help			-	Staff
Research Assistant	-		-	Undergraduate
Postdoctoral Fellow	-			Postgraduate

Second Iteration:

Job Title	**Preceding Class**	**Concurrent Class**	**Succeeding Class**	**New Prediction**
Student Help			Undergraduate	Undergraduate
Research Assistant	Staff		Postgraduate	Graduate
Postdoctoral Fellow	Undergraduate			Postgraduate

Third Iteration:

Job Title	**Preceding Class**	**Concurrent Class**	**Succeeding Class**	**New Prediction**
Student Help			Graduate	Undergraduate
Research Assistant	Undergraduate		Postgraduate	Graduate
Postdoctoral Fellow	Graduate			Postgraduate

In this example, “postdoctoral fellow” is pivotal. Because the job title is so informative, its predicted class during the second iteration is not affected by the wrong prediction for the preceding title (i.e., it is unlikely to transition directly from undergraduate to postgraduate, but it is even more unlikely for a non-postgraduate student to have a job title “postdoctoral fellow”).

Of course, the time gap between the consecutive titles should also be taken into account. For the above example, if the time gap between “research assistant” and “postdoctoral fellow” is more than several years, the initial prediction of undergraduate for the job title “research assistant” may be more appropriate than the revised prediction of graduate.

Here, we do not leverage the time gap between job titles in the model, but do include age, which contains somewhat similar information regarding the timing of job titles.

One issue with the iterated prediction procedure is that the convergence is not guaranteed, especially when there is no pivotal job title. For example, consider the following individual who held two job titles simultaneously.

Table A7 – Illustration of iterative procedure to use concurrent titles

First Iteration:

Job Title	**Preceding Class**	**Concurrent Class**	**Succeeding Class**	**Prediction**
Tutor		-		Graduate
Grader		-		Undergraduate

Second Iteration:

Job Title	**Preceding Class**	**Concurrent Class**	**Succeeding Class**	**Prediction**
Tutor		Undergraduate		Undergraduate
Grader		Graduate		Graduate

Third Iteration:

Job Title	**Preceding Class**	**Concurrent Class**	**Succeeding Class**	**Prediction**
Tutor		Graduate		Graduate
Grader		Undergraduate		Undergraduate

Because we cannot say a tutor or a grader is definitely an undergraduate or graduate, it is possible that, when making a revised prediction, the random forest classifier will simply adopt the classification for the concurrent title predicted in the previous iteration. As a result, the prediction will flip-flop and the algorithm will never stop. Of course, the presence of other people in these occupations mitigates this problem at least to some extent.

<u>Data Construction</u>

As a first step toward incorporating transitional and concurrent classes in the random forest classifiers, we included the manually classified transitional and concurrent classes in our training data in the model rather than predicted occupations. The resulting predictive accuracy is expected to provide an upper bound for the accuracy obtained from the iterated procedure described above.

To construct our sample, the monthly transaction records were collapsed at the individual-title-year level. That is, for each individual, for each calendar year, for each job title held during the year, we kept the individual-title-year record if the individual appeared in the transaction record both before July 1 and after September 30 with the job title. This is to avoid potential noise in the data when annual income is merged. Suppose an undergraduate student held a research assistant position during January through June. Then he or she graduated and obtained a full-time job. If the individual was included in our sample, it would appear that the annual income of the individual is too high to be an undergraduate, and it can potentially mislead the random forest classifier.

The concurrent job titles are defined to be a group of job titles that were held by an individual within a year. When there were multiple concurrent job titles, we selected the one for which the individual was paid for the longest. The preceding job title is defined to be a job title held by an individual in the years preceding the current year. When there were multiple preceding job titles, we selected the most recent one. The succeeding job title is defined to be a job title held by an individual in the years succeeding the current year. When there were multiple succeeding job titles, we selected the one that immediately followed the current job title. Because we restricted our sample to individuals appearing in the transaction data between 2012 and 2014, the occurrence of multiple concurrent or transitional job titles were rare.

Before fitting the random forest classifier, transitional and concurrent classes were binarized because the random forest classifier cannot process categorical data. Each of preceding, concurrent, and succeeding class variables was decomposed into five indicator variables (faculty, postgraduate, graduate, undergraduate, and staff / other).

Methodology

To properly measure the effect of including transitional/concurrent classes on the predictive accuracy, we created the following subsets of observations:

- Everyone: Every observation
- None: Observation without any transitional or concurrent titles
- Prec: Observations with preceding title (may or may not have succeeding or concurrent titles)
- Succ: Observations with succeeding title (may or may not have preceding or concurrent titles)
- Conc: Observations with concurrent title (may or may not have preceding or succeeding titles)
- Any: Observations with at least one of preceding, succeeding, or concurrent title (can have multiple)

We expect that inclusion of transitional/concurrent classes have no effect on occupations where no observations have any transitional/concurrent classes while it will have the largest effect on occupations with many cases with all of the three classes. Each of the six subsets listed above served as a test set, and the random forest classifier was fitted with and without transitional/concurrent classes as predictors.

Regarding the training set, it is unclear whether the set should be restricted in the same way as the test set. Consider the test set “Prec”. On the one hand, it seems reasonable to restrict the training set to only observations with preceding titles. This is because if observations without preceding title were to be included in the training set, the importance of the preceding class in predicting the current class may be discounted. On the other hand, requiring observations in the training set to have preceding titles greatly reduces the number of qualified observations, possibly leading to overfitting. Since the effect of restricting the training set is unclear, we fitted the random forest classifier with and without restriction on the training set.

As shown in Table A8, including the concurrent and transitional occupation has minimal effect on the predictive accuracy. This is most likely due to the limited number of

relevant observations in the training set, and therefore, as more universities participate in the IRIS project and provide data over a longer time period, the concurrent and transitional occupation may become a useful predictor.

Table A8

Features	Training Set	Everyone	None	Prec	Succ	Conc	Any
Job title	Unrestricted	0.83	0.83	0.82	0.84	0.83	0.83
Job title and occupation	Unrestricted	0.82	0.82	0.83	0.85	0.82	0.84
Job title	Restricted to relevant group	0.83	0.82	0.83	0.82	0.86	0.83
Job title and occupation	Restricted to relevant group	0.82	0.83	0.83	0.85	0.78	0.83