Machine learning and domain knowledge

Prasanna (Sonny) Tambe*
The Wharton School, U. Pennsylvania

May 31, 2018

DRAFT COPY, PRELIMINARY AND INCOMPLETE.
Please do not circulate, quote, or post.

Abstract

This paper analyzes a database of online job listings to investigate the structure of machine learning jobs. It first presents descriptive statistics about the demand side of the machine learning labor market. Second, it tests the hypothesis that unlike prior waves of IT, machine learning skills tend to be bundled with domain knowledge, such as economics or life sciences. Complementarities between technical skills and domain knowledge are also reflected in the college degrees listed in job ads. When compared with more traditional technical jobs, job postings related to machine learning often indicate that applicants can either have degrees in technical disciplines or subject areas related to the data context. In general, the findings indicate that the adoption of machine learning skills into jobs is likely to be governed by the need to couple these skills with domain knowledge, potentially turning a wider category of jobs into "technical" jobs. Implications for labor market competition and education policy are discussed.

Introduction

The potential for AI and machine learning to transform the labor force has been attracting widespread attention. Early academic work in this area principally focuses on how automation and AI technologies might substitute for workers (Acemoğlu and Restrepo, 2016; Brynjolfsson et al., 2018). There has also been recent interest in the question of where prediction tasks complement human capital (Agrawal et al., 2016; Choudhury et al., 2018). This paper leans towards the second question, and analyzes job skills for workers who are expected to build and use these new AI-based systems. Specifically, it argues that machine learning (ML) related skills fundamentally differ from prior generations of technical skills because domain knowledge and ML skills are complements and are difficult to separate into different jobs. Therefore, jobs that require machine learning skills also often

*Correspondence: tambe@wharton.upenn.edu. I am grateful to Dan Restuccia and Bledi Taska for providing access to the Burning Glass Technologies data.

require subject matter expertise in fields such as economics or microbiology. As a result, machine learning is likely to be integrated into jobs outside of traditional technical functions.

This is important because many stakeholders have been grappling with the question of how machine learning should be integrated with the requisite domain knowledge needed for machine learning tasks. If data analysis skills such as machine learning tasks diffuse into non-IT jobs to a greater extent than prior waves of technological innovation, it has implications for the structure of the labor market. A number of papers have focused on how the unique economic characteristics of technical skills, such as their portability and rapid depreciation rates, affect the structure of technical labor (Saxenian, 1996; Fallick et al., 2006; Tambe et al., 2018). Therefore, if machine learning bleeds into a new class of occupations, it can bring some of market features to these occupations. This question also has implications for education. Greater emphasis is being placed on coding in basic education, due to a belief held by many that coding will be a basic and required skill. An open question is whether employers will separate technical tasks from other functions, as they have done in the past with technologies such as database management, or whether machine learning methods and the required domain knowledge will be coupled together in a way that leads to more hybrid jobs in the future. The latter has practical implications–such as whether MBA students need machine learning skills– and may imply an important supply-side constraint to adjustment to the AI economy, because it may be challenging to produce workers who can simultaneously bring both technical skills and domain skills to a job.

To the best of my knowledge, no prior work has addressed this question. Research on the effects of AI on the labor market has been rapidly evolving, but there has been limited empirical analysis on how technologies change the skills required within jobs, or more generally, how jobs differ across firms with different production functions. Empirically, this is a challenging topic for two reasons. First, this paper focuses on *intra-occupational* task shifts, which are very difficult to measure using administrative data. Although revisions to the O-NET data capture skill-based changes within occupations that occur over time, they do not track how differences in production choices are associated with differences in job skills, especially at the level of individual technologies. In other words, how the skills required to be a "marketing analyst" at leading tech employer differ from the skills required for the same job title at an “old economy firm”. Yet, by all accounts, this margin of change has been an important one for job transformation. Prior work makes related arguments

about the relative importance of the intensive and extensive margin of changes to job skills (Autor et al., 2003). Second, penetration levels of AI technologies into the US commercial sector remain limited and the data sources that have traditionally been used to analyze changes in the skill content of the US workforce are available only at the industry level. Analyses based on such data often require greater penetration levels to produce detectable effects of changes in labor composition. Moreover, skill based differences within occupations, such as growing demand for deep learning skills within software job classifications or greater demand for data visualization within marketing analyst occupations, are normally invisible in these data.

This paper leverages data on online job listings to analyze how the adoption of machine learning technologies is impacting job structure. The data provider aggregates labor market information from over 17,000 independent job boards and corporate web sites to develop a repository of US job vacancy listings. In each of the seventy two months that span the period analyzed in this paper (2010-2016), there are approximately one and a half million different job listings for a total of almost eighty million job listings in the data set. A benefit of the database is that the provider has processed these data to map job vacancies to location, occupation, and industry codes and they have parsed and normalized the job title and skill data. Notably, this database includes information on differences in skill requirements within listings for the same job title. In this analysis, these job listings data are essentially treated as survey data where employers are asked "What skills would you prefer for someone to work in this job?".

Job listings data have been used extensively in prior research because of the advantages listed above—they enable detailed measurement of changes in what firms want. For this reason, they have been an especially popular source of data to study changes in the IT labor market (Slaughter and Ang, 1996; Gallivan et al., 2002). With respect to the empirical challenges associated with using administrative data, analysis of job vacancies provide two key advantages. The first is the granularity of the skills data in each job listing, which allows measurement of whether specific occupations—for example, risk analysts—are increasingly likely to require familiarity with tools such as neural nets or supervised learning. The second benefit of these data is that job listings are an early indicator of organizational skill-based changes. Job listings report a gradient in terms of where the workforce is heading. Although it may take years for skill-based changes to produce statistically detectable differences at the organizational level, changes to human capital policies are

fairly immediate (see Bresnahan et al. 2002 for a similar argument in the context of human capital screening and investment policies). Online job listings also indicate the skills that organizations are searching for, not the outcome of the job search process. If the labor force is unable to supply the skills required by the organization, employment data will not reflect employers' needs until some adjustment has occurred. This study is principally interested in how organizations are expressing changes their short-run needs, which are more rapidly reflected in employer job listings than by labor force skills. They indicate where the firm is headed.

Of course, there are important caveats associated with the analysis of these data. The largest obstacle to the analysis of these data is that the sampling frame is not well defined. The online job vacancy data do not tell us anything about positions that employers choose not to post online but instead, fill through offline channels such as print advertising or word-of-mouth. Jobs that employers post online and offline can also differ in systematic ways. Second, care is required in interpreting these listings as an economic quantity. An online job posting may not indicate a job vacancy. Employers can be slow to update online listings so an online listing can persist long after a position has been filled. The research question chosen in this paper is meant to minimize as many of these challenges as possible since the listings are not treated as measures of vacancies or demand, but the unknown sampling frame still may introduce sources of bias that are discussed later in the paper.

Key findings of the paper are that machine learning skills are much more distributed among jobs than prior generations of technical skills. 78% of jobs that use programming languages like Java and C++ are IT jobs. By comparison, only 34% of jobs that use machine learning skills are IT jobs. Another 30% of these are jobs related to economic analysis or strategic planning or life sciences research. This finding suggests that it is easier to integrate the machine learning into these jobs than it is to introduce the necessary domain expertise into IT jobs. Moreover, conditional on both job title and industry, machine learning skills in jobs increase the likelihood that the job also requires domain knowledge, relative to a similar job that requires technical skills like C++ or Java, but does not require machine learning skills, per se. There is supporting evidence from the educational degrees that the job listings recommend that the applicants have. Relative to job listings with the same title and industry that require knowledge of other technical skills, jobs that need machine learning skills are more likely to state that applicants can have degrees either in technical disciplines or in

subject-specific domains like finance or life sciences.

This study adds to the discussion on how the growing penetration of data-driven and machine learning practices within industries will affect job skills (e.g. see Freeland 2012) and it is embedded in a literature on the effects of IT-enabled change on the skill composition of the workforce (Bresnahan et al., 2002; Autor et al., 2003). Although this literature has considered the effects of IT on skill and task demand, an important caveat of analyses such as these is that they require fixing technological capabilities at a particular point in time. Prior work has principally considered IT as technology that lowers communication costs (e.g. networked technologies), but the effects of IT for facilitating advanced data analysis has not yet been studied, and has the potential to alter our current understanding of job structure.

Theory and background

The literature on the effects of IT investment on labor demand argues that the diffusion of IT over the last several decades has led to "skill-biased technical change", as the adoption of information technologies has increased demand for high-skill workers, most often identified as workers with a college education. College-educated workers complement IT-enabled production methods because IT lowers communication costs, which makes it easier to decentralize decisions, and college educated workers are better equipped for production characterized by decentralized decision-making, team-based production, critical thinking, and problem solving. Empirical evidence for these claims has been provided in a variety of contexts (Bresnahan et al., 2002; Caroli and Van Reenen, 2001). Papers in this literature have also developed a deeper characterization of the skills for which IT complements or substitutes (Levy and Murnane, 1996). By using occupational skill data from the BLS O-NET, Autor and colleagues argue that IT substitutes for routine mental and cognitive skills and complements non-routine mental and cognitive skills (2003). They argue that this characterization explains much IT-related change in skills.

Much of the emphasis in this literature has been placed on the distinction between routine and non-routine work. The former of these requires workers to follow rules and the latter requires them to handle exceptions. However, what can be deemed routine has been changing, and IT has advanced in a way such that jobs that were unaffected by the advance of IT just a few years ago are now

being impacted by newer IT advances, sometimes in unpredictable ways. One example is driving an automobile, which was once deemed beyond the reach of automation but is now, arguably, a target for wide scale automation. As a result, there have been recent attempts to update older and static characterizations of the extent to which IT can complement or substitute for various tasks. In particular, technological progress since the 2000's has focused on networked computing, data collection, storage, and most recently, the leveraging of these growing data assets through higher order analysis that can drive rapid and effective decisions.

The growing importance of data analysis within firms has the potential to further alter what skills are required from workers in an economy where AI plays an important role. The theoretical underpinnings of this argument are drawn from a literature on information and decision-making within organizations. Perhaps to a larger extent, however, it is informed by an appreciation of what modern data technologies do. Modern data analysis technologies facilitate the capture, processing, retrieval, and analysis of large data sets. When working with prior waves of information technology, the domain and decision context has largely been separable from workers building IT systems. However, machine learning and data science tasks often require an understanding of modeling choices as well as the data and how it was collected and what the different measured fields mean to the organization. When missing the technical skills or the domain context, it becomes difficult to articulate what questions can be asked of the data. More generally, the rising importance of data assets within organizations raises the returns not only to retrieving the data (data management and manipulation), but also analyzing it (data analysis), and presenting it (data visualization and articulation and presentation). Unlike the management, manipulation, and reporting of data, the latter two categories (data visualization and presentation) require knowledge of the domain to be effective. The combination of data and domain skills has been limited to a few niche occupations, such as actuarial science, but this paper argues that this combination of skills will become more widespread if machine learning technologies continue to spread.

Job listings data

Description

This study analyzes a database of online job listings. Job listings have informed a number of earlier studies on the changing skill requirements of the technical workforce (Todd et al., 1995; Slaughter and Ang, 1996; Gallivan et al., 2002; Lee and Han, 2008; Litecky et al., 2008; Kennan et al., 2008; Mathews and Pardue, 2009). A noteworthy advantage of job listings data for analyses like these is that they capture granular changes in the demand for skills, such as changes in the popularity of specific IT skills (e.g. COBOL, C++, Java) at a level that is not tracked by administrative data sources.

One difference between the database analyzed in this study and the job listings data used in prior studies is the scale of the data source. Most prior work on job listings has relied on hand-collected samples, but this study benefits from an industry of real-time labor market information aggregators that 1) uses software to crawl online data at "Internet scale" and 2) uses natural language processing tools to parse skills and other job information, which would otherwise require a prohibitive amount of labor due to the extremely large size of the corpus. The number of job ads analyzed in prior work has been small; for the studies listed above, 2,108 (Gallivan et al., 2002), 1,234 (Todd et al., 1995), 906 (Mathews and Pardue, 2009), and 1,560 (Slaughter and Ang, 1996). By comparison, this database includes information for over 80 million new job ads posted over 72 months. Sample size is not necessarily an advantage beyond providing greater statistical power, but the size and scope of the data are useful for this analysis because they are simultaneously comprehensive and expansive, which allows generalized comparisons to be made across labor markets and occupations, and because they are granular, which enables analyses at the level of detailed skills (e.g. neural networks or marketing research).

To generate the database, the data provider uses proprietary software to collect and parse data from over 17,000 independent job boards and corporate web sites. The data include job posting date, job location, employer name, job title, required degrees and certifications, and normalized skills required for the job. The skill data are normalized using a taxonomy created and maintained by the data provider. All of these data are extracted directly from the job listing using natural language processing tools and duplicates in the data are removed. In addition, the job title in each

record is tagged with an occupational code (O-Net) and the employer for each listing is associated with an industry (NAICS). Of course, there are also several challenges when analyzing data like these. These are both discussed further below, after the main specification is introduced.

Empirical framework

Key measures

The key independent variable used in this analysis is whether a job listing requires skills related to machine learning. The list of skills used to identify machine learning is:

Table 1: Machine learning skills

Machine Learning	Artificial Intelligence
Predictive Modeling	Predictive Analytics
Predictive Models	Data Mining
Deep Learning	Neural Networks
K-Means	Cluster Analysis
Natural Language Processing	

There is some measurement error if jobs are misidentified as requiring machine learning skills (unlikely) or if they are not identified as requiring machine learning skills but they are missed in the list above.

The key dependent variables are the presence of various families of skills used in job listings, including domain skills, creativity, and so on. These measures are binary, indicating whether or not a skill from a skill family appears in a listing. The skills used to identify each skill family are shown in Appendix A.

Regression model

The analysis tests whether certain skills are more likely to appear in listings for jobs that use machine learning. The argument is that growing volumes of data have simultaneously raised demand for better data analysis technologies, data-driven organizational practices, and the skills required to convert data into decisions. The underlying shift in the volumes of information available to the firm, facilitated by Moore's Law and cheaper instrumentation, have driven the demand for the IT

innovation and skill-biased changes observed in the data. The regression model estimates whether skills are more likely to co-appear with machine learning technologies. The model can be written:

$$SKILL_i = ML_i + \gamma_i + \rho_i + t_i + \epsilon_i \tag{1}$$

The unit of observation is the job listing. The dependent variable is an indicator of whether or not listing i lists a skill (e.g. *Economics*, *Predictive Modeling*). The key independent variable is whether or not the vacancy uses machine learning technologies (one of the skills in Table 1). Fixed effects are included for the job title, the employer's industry and in many specifications, the employer and the geographic region in which the vacancy is located. In the most robust specification, this regression tests whether, when comparing listings posted by the same employer, in the same region, and for the same job title, the presence of machine learning skills is associated with specific skill families.

Sources of bias

Omitted variables

The specifications used in this paper include both job title and industry fixed effects. In other words, the variation comes from "software developers" or "marketing research analysts" within the same industry and firm that differ in terms of whether or not the job postings require skills related to machine learning. To further minimize heterogeneity within workers in the same job title, the control group is limited based on skills. Comparison workers are those workers that not only share the same job title, region, and employer, but that also have a foundation of software skills that indicates a certain level of technical expertise (e.g. workers with expertise in C++, Unix, Java). The error term ϵ_i represents the effect of any other unobserved factors that might otherwise influence the value of the dependent variable (i.e. whether or not a particular skill is posted in an ad).

Sampling

As discussed above, the analysis of Internet data can be challenging because the sampling frame is not clear. This study bypasses some of the issues that might be associated with analyzing online job listings because it uses the data to measure the skill content of jobs, not as a measure of job vacancies

or quantities of labor demanded. In other words, the data are treated like a survey administered to a large sample of employers asking, "What skills would be ideal for an applicant to have to do the job listed in the ad?" Given this context, the population of interest is jobs in organizations, not job vacancies or job openings. Therefore, the results should be interpreted as reflecting the preferences of employers that post job ads online, which may be differ from the average employer in systematic ways.

Results

Descriptive statistics on the machine learning workforce

This section presents some simple descriptive statistics, drawn from the job listings data, about the demand side of the market for machine learning skills. Figure 1 illustrates the growth in popularity of machine learning skills in job listings over the last decade. There is a steady month-on-month growth in job listings requiring machine learning. Some common job titles that use machine learning skills are shown in Table 2. These titles are the level of granularity at which the fixed-effects are included in the key specification. They key comparisons in the sections below are between job ads for jobs such as these where machine learning is included in the list of skills and ads for which they are not.

When compared with the distribution of all vacancies from the underlying sample, the vacancies related to modern data technologies are concentrated in information industries and professional and scientific and technical services, which is not surprising because these industries include many prominent technology companies, and are in general, information intensive. However, there is also significant big data hiring in finance and retail trade industries, which is consistent with the purported prevalence of data analysis practices in retail firms such as Walmart and in many large banks and insurance companies.

Table 3 shows the distribution of machine learning skills and other technical skills by career category of the job ad. The key takeaway from Table 3 is that machine learning skills are spread over different career groupings. Only about 34% of machine learning skills are in IT jobs, compared with over 80% of jobs for technical skills such as C+ and Java. This provides some early evidence in support of the notion that machine learning tasks are likely to be spread over a larger set of

career categories than traditional technical skills. A key goal of the paper is to test the hypothesis that is due to the need to couple machine learning skills with the domain knowledge necessary to understand the data.

Table 4 shows summary statistics for some of the key skills analyzed in the data. About 6% of the ads in the sample list machine learning skills. By comparison, coding shows up in 40% of the job ads in the sample. About 14.5% of job ads include some type of domain skill.

Machine learning and other technical skills

This section tests for correlations between machine learning skills and other technical tasks, partly as a check on consistency in the data. Table 5 reports these results. Machine learning skills are negatively correlated with general database skills (t=-1.86). They are positively correlated with coding skills (t=6.55). Machine learning skills are also positively associated with experiments (t=6.33) and statistics (t=29.29). They are also positively associated with technical skills commonly paired with data science and learning: Python, R, and Apache Hadoop.

Machine learning and higher order skills

Table 6 reports results from regressions of machine learning skills on higher order skills. Machine learning skills in listings are positively correlated with creativity (t=1.77). It is also a strong predictor of online research skills (t=5.24). The ability to analyze and present data is also very important for employees who are expected to use machine learning on-the-job. The coefficient estimates on visualization (t=5.25), analytics (t=16.5), and decision-making are all positive and statistically significant (t=2.17).

Machine learning and domain knowledge

Table 7 reports regressions of machine learning skills on domain knowledge, where domain knowledge in this table is measured as an aggregated measure including knowledge from a number of different disciplines (e.g. science, health, finance, accounting, statistics). These are the main results of the paper. Like in the earlier tables, all of the regressions include job title and industry fixed-effects, so the variation in the regression is between workers in the same job title and industry, but across groups of jobs where some require machine learning and some do not.

The coefficient estimates on the ML measure in column (1) indicate that ML skills are positively associated with the presence of domain skills in the listing (t=16.0). As a falsification test, this table also includes regressions of other families of technical skills. The second column analyzes jobs that list technical skills related to front-end web development. The coefficient estimates from this regression suggest that a negative correlation exists between job ads searching for front end web development skills and those that are searching for domain skills (t=-6.25). The third column uses the LAMP technology stack (Linux, Apache, MySQL, PHP) which is a open source task that is used to build web sites. The coefficient estimate for this regression is not significantly different than zero. For database administration skills, the coefficient estimates are also negative and significant (t=5.57).

These effects are robust to including all indicators for all four sets of technical skills within the same regression. The coefficient estimate on ML skills are the only ones that are positively correlated with domain skills (t=15.87). These correlations are all robust to including both regional and employer fixed effects in columns (6) and (7).

Table 8 uses more precise measures of domain skills. The coefficient estimates are positive and significant for a number of specific domain skills in the data, including biology (t=6.5), business strategy (t=11.8), economics (t=2.57), marketing (t=4.67), and statistics (t=8.0).

Machine learning and college degrees

If ML-related job openings require different forms of expertise, this should also be reflected in the educational degrees listed in the job ads. Table 9 reports results from regressions of machine learning skills on the educational degrees listed in the job ad. The first column is a regression of ML-related skills on whether the listing requests a computer science or information systems (CS/IS) degree. The coefficient estimate on the ML-related skills measure suggests that CS/IS degrees are less frequently listed for machine learning skills than for other technical skills (t=-8.75).

On the other hand, ML-related skills are positively associated with the ad listing college degrees in science (t=5.50), business (t=6.67), stats (t=5.67), or economics (t=5.67) degrees as useful background for the job. For column (6), the dependent variable is marked as a 1 if the ad lists degrees from a technical discipline as well as degrees in a domain as potential options for applicants. Consistent with this notion, the coefficient estimate on ML-related skills being listed in the job ad increase the likelihood that the job listing lists both domain and technical degrees as potential

options for candidates to have (t=6.5).

Conclusions

Using a new source of data on the skills sought by employers when filling jobs, this analysis provides evidence that jobs requiring machine learning skills also require greater levels of domain knowledge. These findings provide a window into the labor market changes that are occurring due to the spread of machine learning technologies in the US economy. Just as prior waves of IT-enabled change have impacted skill levels for workers in all parts of the organization, data driven technologies have the potential to further change the skill requirements for workers who interact with the growing volumes of data collected and used by organizations to make decisions.

This study suggests a number of avenues for future research. The posting of online job listings can differ from the total labor demanded by firms in ways that are hard to track through online listings. A significant limitation of this study, especially from a policy perspective, is that this paper analyzes changes in the skills required by employers, but cannot say anything about the supply-side of these skills, or about how effective existing institutions or educational systems will be for providing these skills to current or future workers. These issues notwithstanding, this paper provides evidence of the direction in which organizations are moving. Finally, although the data source has some information on wages, there is limited information on the quality of these new jobs. This paper reserves judgment on whether these labor market changes are good or bad for workers.

There is significant scope for future work in this area. By most accounts, we are at the beginning of a large wave of investment in technologies that aid in converting data to decisions. Of course, the nature of technological capabilities continues to evolve. New technologies associated with data collection, analysis, prediction, and visualization will have improved capabilities to generate insights. As this boundary is pushed forward, it will continue to change which skills are required by employers and how they are distributed within the firm.

[1] Related work analyzes offshoring using a similar framework, which, although enabled by the lower communication costs afforded by IT, can affect labor demand in ways that differ from IT. This literature focuses on the costs associated with codifying and shipping the inputs and outputs required for tasks across distances and across firm boundaries. A principal finding from this literature is that "" tasks, which require face-to-face contact, are less likely to be offshored (Blinder, 2009; Tambe and Hitt, 2012).

[2] See `https://onet.rti.org/survey.cfm` for further details on the O-Net data collection process.

Figure 1: Growth in job listings requiring machine learning skills

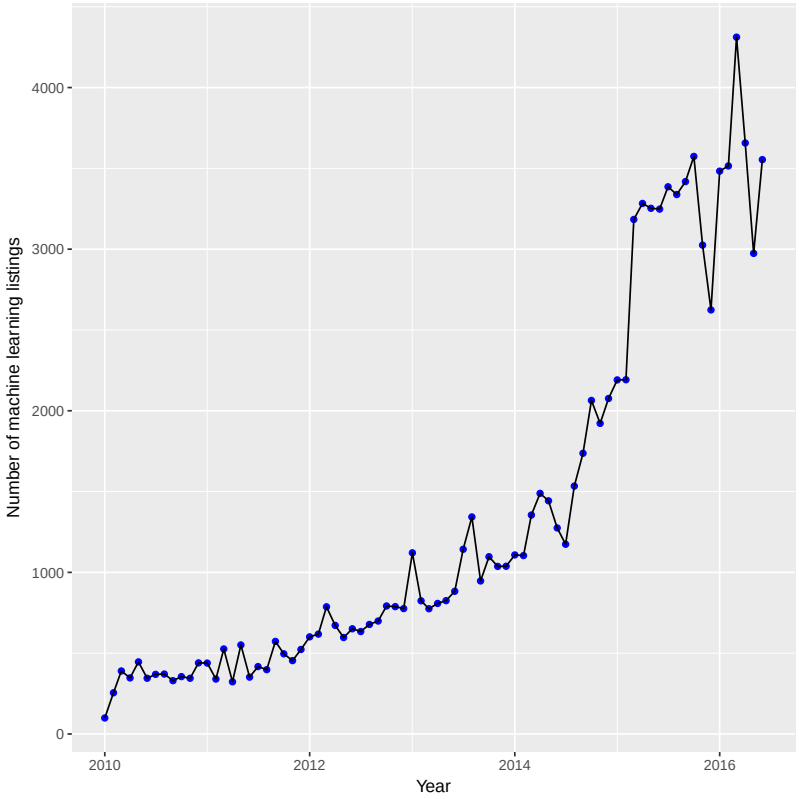


Table 2: Sample common job titles for postings require machine learning skills

Software Engineer
Senior Software Engineer
Business Analyst
Financial Analyst
Account Manager
Web Developer
Java Developer
Recruiter
.Net Developer
Project Manager

Table 3: Machine learning skills by career category

ML	% (ML)	Other	% (Other)
Information Technology	34.4	Information Technology	82.4
Planning and Analysis	28.6	Planning and Analysis	2.8
Finance	5.9	Engineering	2.4
Marketing and Public Relations	4.5	Sales	1.4
Sales	3.1	Business Management and Operations	0.9
Business Management and Operations	2.7	Finance	0.8
Engineering	2.6	Marketing and Public Relations	0.6
Health Care including Nursing	1.5	Manufacturing and Production	0.5
Human Resources	1.5	Design, Media, and Writing	0.4
Science and Research	1.2	Human Resources	0.3

Numbers in table denote percentages of skill in that career category.

Table 4: Summary stats for machine learning and other related skills

Statistic	N	Mean	St. Dev.	Min	Max
ML	11,008	0.062	0.242	0	1
Data	11,008	0.461	0.498	0	1
Coding	11,008	0.397	0.489	0	1
Experiments	11,008	0.006	0.079	0	1
Statistics	11,008	0.027	0.161	0	1
Domain	11,008	0.145	0.352	0	1
Python	11,008	0.210	0.407	0	1
R	11,008	0.015	0.123	0	1
Hadoop	11,008	0.043	0.203	0	1
No. of skills	11,008	16.078	9.017	1	71

Table 5: Technical skills related to machine learning

	Dependent variable:						
	Data	Coding	Experiments	Statistics	Python	R	Hadoop
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ML	-0.041^{*}	0.131^{***}	0.019^{***}	0.205^{***}	0.259^{***}	0.204^{***}	0.158^{***}
	(0.022)	(0.020)	(0.003)	(0.007)	(0.018)	(0.005)	(0.009)
Log(No. of skills)	0.251^{***}	0.090^{***}	0.002^{**}	0.019^{***}	0.065^{***}	0.014^{***}	0.035^{***}
	(0.006)	(0.006)	(0.001)	(0.002)	(0.005)	(0.002)	(0.003)
Constant	0.155^{***}	-0.116^{***}	-0.003	-0.049^{***}	-0.154^{***}	-0.038^{***}	-0.083^{***}
	(0.019)	(0.018)	(0.003)	(0.006)	(0.016)	(0.005)	(0.009)
Observations	10,765	10,765	10,765	10,765	10,765	10,765	10,765
Adjusted R^2	0.228	0.310	0.197	0.142	0.192	0.153	0.079

Note: *p<0.1; **p<0.05; ***p<0.01

All regressions include job title and industry fixed effects.

Table 6: Higher order skills complementary to machine learning

	Dependent variable:				
	Creativity	Research	Visualization	Analytics	Decision-Making
	(1)	(2)	(3)	(4)	(5)
ML	0.023*	0.089***	0.021***	0.132***	0.013**
	(0.013)	(0.017)	(0.004)	(0.008)	(0.006)
Log(No. of skills)	0.039***	0.073***	0.004***	0.012***	0.010***
	(0.003)	(0.004)	(0.001)	(0.002)	(0.001)
Constant	−0.057***	−0.128***	−0.006*	−0.015**	−0.023***
	(0.010)	(0.013)	(0.003)	(0.006)	(0.004)
Observations	9,768	9,768	9,768	9,768	9,768
Adjusted R^2	0.054	0.073	0.017	0.093	0.054

Note: *p<0.1; **p<0.05; ***p<0.01
All regressions include job title and industry fixed effects.

Table 7: Machine learning and domain knowledge

	Dependent variable:						
	Domain knowledge required for job					MSA effects	Employer effects
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ML	0.240***				0.238***	0.246***	0.308***
	(0.015)				(0.015)	(0.015)	(0.033)
Front end		−0.050***			−0.051***	−0.054***	−0.029*
		(0.008)			(0.008)	(0.008)	(0.017)
LAMP			0.005		0.010	0.024**	0.041*
			(0.010)		(0.010)	(0.010)	(0.023)
DBA				−0.039***	−0.038***	−0.032***	−0.032**
				(0.007)	(0.007)	(0.007)	(0.014)
Log(No. of skills)	0.075***	0.084***	0.077***	0.088***	0.090***	0.090***	0.092***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.011)
Constant	−0.134***	−0.144***	−0.140***	−0.136***	−0.130***	−0.220	−0.013
	(0.013)	(0.013)	(0.014)	(0.013)	(0.013)	(0.153)	(0.218)
Observations	10,881	10,881	10,881	10,881	10,881	10,881	4,551
Adjusted R^2	0.175	0.159	0.156	0.159	0.180	0.214	0.629

Note: *p<0.1; **p<0.05; ***p<0.01
All regressions include job title and industry fixed-effects.

Table 8: Machine learning and examples of domain knowledge

	Dependent variable:						
	Biology/Genetics	Business Strategy	Economics	Marketing	Mathematics	Optimization	Statistics
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ML	0.013***	0.059***	0.018**	0.014***	0.062***	0.068***	0.024***
	(0.002)	(0.005)	(0.007)	(0.003)	(0.013)	(0.011)	(0.003)
Log(No. of skills)	−0.001	0.007***	0.008***	0.001	0.023***	0.028***	0.0002
	(0.0004)	(0.001)	(0.002)	(0.001)	(0.003)	(0.003)	(0.001)
Constant	0.002*	−0.017***	−0.013**	−0.002	−0.034***	−0.047***	0.001
	(0.001)	(0.004)	(0.005)	(0.002)	(0.010)	(0.008)	(0.002)
Observations	10,702	10,702	10,702	10,702	10,702	10,702	10,702
Adjusted R^2	0.237	0.218	0.105	0.045	0.037	0.024	0.016

Note: *p<0.1; **p<0.05; ***p<0.01
All regressions include job title and industry fixed-effects.

Table 9: ML-related job listings and preferred college degrees

	Dependent variable:					
	CS/IS	Science	Business	Stats/OR	Economics	Dual
	(1)	(2)	(3)	(4)	(5)	(6)
ML	-0.105^{***}	0.022^{***}	0.060^{***}	0.017^{***}	0.017^{***}	0.104^{***}
	(0.012)	(0.004)	(0.009)	(0.003)	(0.003)	(0.016)
Log(No. of skills)	-0.001	0.00000	0.004	0.0003	-0.004^{***}	0.018^{**}
	(0.005)	(0.002)	(0.004)	(0.001)	(0.001)	(0.007)
Constant	0.984^{***}	0.002	-0.006	-0.001	0.010^{**}	-0.029
	(0.016)	(0.006)	(0.013)	(0.005)	(0.004)	(0.022)
Observations	4,098	4,098	4,098	4,098	4,098	4,098
Adjusted R^2	0.415	0.114	0.520	0.057	0.022	0.072

Note: *p<0.1; **p<0.05; ***p<0.01
All regressions include job title and industry fixed-effects.

References

Acemoğlu, D. and Restrepo, P. (2016). The race between machines and humans: Implications for growth, factor shares and jobs.

Agrawal, A., Gans, J., and Goldfarb, A. (2016). Exploring the impact of artificial intelligence: Prediction versus judgment. *University of Toronto*.

Autor, D., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4).

Autor, D. and Price, B. (2013). The changing task composition of the US labor market: An update of Autor, Levy, and Murnane (2003).

Autor, D. H., Levy, F., and Murnane, R. J. (2002). Upstairs, downstairs: computers and skills on two floors of a large bank. *ILR Review*, 55(3):432–447.

Blinder, A. S. (2009). How many US jobs might be offshorable? *World Economics*, 10(2):41.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics*, 117(1).

Brynjolfsson, E., Mitchell, T., and Rock, D. (2018). What can machines learn and what does it mean for occupations and industries? In *AEA Papers and Proceedings*.

Caroli, E. and Van Reenen, J. (2001). Skill-biased organizational change? evidence from a panel of British and French establishments. *Quarterly Journal of Economics*, pages 1449–1492.

Choudhury, P., Agarwal, R., and Starr, E. (2018). Different strokes for different folks: Experimental evidence on complementarities between human capital and machine learning.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Fallick, B., Fleischman, C. A., and Rebitzer, J. B. (2006). Job-hopping in silicon valley: some evidence concerning the microfoundations of a high-technology cluster. *The Review of Economics and Statistics*, 88(3):472–481.

Freeland, C. (2012). In big data, potential for big division. *New York Times*, (January 12).

Gallivan, M., Truex III, D. P., and Kvasny, L. (2002). An analysis of the changing demand patterns for information technology professionals. In *Proceedings of the 2002 ACM SIGCPR conference on Computer personnel research*, pages 1–13. ACM.

Horton, J. and Tambe, P. (2015). Labor economists get their microscope: Big data and labor market analysis. *Big Data*, 3(3):1–8.

Kennan, M. A., Willard, P., Cecez-Kecmanovic, D., and Wilson, C. S. (2008). IS knowledge and skills sought by employers: a content analysis of australian IS early career online job advertisements. *Australasian Journal of Information Systems*, 15(2).

Lee, C. K. and Han, H. (2008). Analysis of skills requirement for entry-level programmer/analysts in Fortune 500 corporations. *Journal of Information Systems Education*, 19(1):17.

Levy, F. and Murnane, R. J. (1996). With what skills are computers a complement? *The American Economic Review*, pages 258–262.

Litecky, C., Prabhakar, B., and Arnett, K. (2008). The size of the IT job market. *Communications of the ACM*, 51(4):107–109.

Mathews, J. M. and Pardue, H. (2009). The presence of IT skill sets in librarian position announcements. *College & Research Libraries*, 70(3):250–257.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.".

Saxenian, A. (1996). *Regional advantage*. Harvard University Press.

Slaughter, S. and Ang, S. (1996). Employment outsourcing in information systems. *Communications of the ACM*, 39(7):47–54.

Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science*, 60(6):1452–1469.

Tambe, P. and Hitt, L. M. (2012). Now IT's personal: Offshoring and the shifting skill composition of the US information technology workforce. *Management Science*, 58(4):678–695.

Tambe, P., Ye, X., and Cappelli, P. (2018). Paying to program? engineering brand and high-tech wages.

Todd, P. A., McKeen, J. D., and Gallupe, R. B. (1995). The evolution of IS job skills: a content analysis of is job advertisements from 1970 to 1990. *MIS Quarterly*, pages 1–27.

Appendix A

Table 1: Domain skills

Skill family	Skills
MARKETING	Merchandising
	Sales Planning
	Retail Management
	Market Research
	Product Knowledge
	Sales Goals
	Description and Demonstration of Products
STRATEGY	Business Strategy
	Market Strategy
	Key Performance Indicators
	Strategic Planning
	Business Planning
SCIENCES	Biochemistry
	Genome
	Molecular Biology
	Chemistry
	Genetics
	Immunology
	Biotechnology
	Microbiology
	Biology
	Biostatistics
	Metabolic Pathways
	Anatomy
ECONOMICS	Economics
	Econometrics

Table 2: Technology stacks

Skill family	Skills
DBA STACK	SQL
	Microsoft SQL
	SAP
	ERP
LAMP STACK	Java
	Linux
	Apache
	MySQL
	PHP
FRONT END STACK	Javascript
	Ruby on Rails
	PHP
	HTML5
	CSS