work2vec: Using Language Models to Understand Wage Premia

Sarah H. Bana*

March 30, 2022

Abstract

Does the text content of a job posting predict the salary offered for the role? There is ample evidence that even within an occupation, a job's skills and tasks affect the job's salary. Capturing this fine-grained information from postings can provide real-time insights on prices of various job characteristics. Using a new dataset from Greenwich.HR with salary information linked to posting data from Burning Glass Technologies, I apply natural language processing (NLP) techniques to build a model that predicts salaries from job posting text. This follows the rich tradition in the economics literature of estimating wage premia for various job characteristics by applying hedonic regression. My model explains 87 percent of the variation in salaries, 26 percent (18 percentage points) over a model with occupation by location fixed effects. I apply this model to the question of online certifications by creating counterfactual postings and estimating the salary differential. I find that there is substantial variation in the predicted value of various certifications. As firms and workers make strategic decisions about their human capital, this information is a crucial input.

*Stanford Digital Economy Lab. Email: sarah.bana@gmail.com. This paper leans heavily on the methods described in work by myself, Erik Brynjolfsson, Daniel Rock, and Sebastian Steffen entitled "job2vec: Learning a Representation of Jobs." Throughout the course of this project, I was told that "job2vec" was trademarked. Accordingly, we have adjusted to using "work2vec." I am very grateful for thoughtful comments from seminar participants at the Stanford Institute for Human-Centered Artificial Intelligence, Boston University, Emory University, Simon Fraser University, the University of Minnesota, the Brynjolfsson Lab, MIT Initiative on the Digital Economy, University of California Santa Barbara Applied Lunch, Chapman University, and the University of Nevada at Reno, along with countless friends and family members. Special thanks for the funding from the Stanford Institute for Human-Centered Artificial Intelligence's Google Cloud Credit Grant that enables this research. All errors are my own.

1 Introduction

It is estimated that 1.7 megabytes of data are created every second for every person on Earth (Domo, 2018). This is largely consistent with more and more of our activities being conducted online. One activity increasingly performed online is job search – firms post jobs and interact with candidates virtually, and workers use online job boards to identify opportunities (Kuhn and Mansour, 2014). The data produced from these activities has the potential to generate unprecedented insights into firms' production functions and workers' activities.

Such data is often unstructured. For example, job postings are free-form text of varying length, without well-defined fields. Thus far, researchers have condensed such text into structured data by identifying relevant key words or adding high dimensional fixed effects by categorizing jobs into discrete buckets.

The text is rich and new tools in computer science have demonstrated breakthrough performance in "understanding" text (Devlin et al., 2018). At the same time, a paradigm shift in artificial intelligence (AI) systems has led to the growth of foundational models – models that are trained on broad data at scale and can be adapted to a wide range of downstream tasks (Bommasani et al., 2021). This approach significantly reduces the computational cost of using text data. One foundational model, Bidirectional Encoder Representations from Transformers (BERT), provides context-dependent embeddings (dense vectors) that can readily be used as the first layer of a model.

With these tools, I train a natural language processing (NLP) model on the text of job postings, and demonstrate that the text of the posting matters. This model takes the text of the posting as an input and translates the text to vectors using BERT's pre-trained word embeddings. These word embeddings, for example, will produce different vectors for the word "models" when characterizing a job advertisement that states "deploy machine learning models" compared to "models exceptional customer service." This initial layer produces a matrix of 512 by 768 dimensions for each job posting. Additional model layers condense the dimensionality.

Using a new dataset with salary information from Greenwich.HR linked to posting data from Burning Glass Technologies, I can reframe salary prediction as a supervised learning problem. My model, incorporating the text, explains 87 percent of the variation in salaries, a 26 percent (18 percentage point) increase over a baseline with occupation fixed effects by Metropolitan Statistical Area (MSA) fixed effects. On another relevant metric, the natural language processing model represents a 42.9 percent decrease in the out-of-sample Root Mean Square Error (RMSE).

The language model outperforms other supervised learning models taking into account the skill clusters tagged in the postings, suggesting that the context matters for salaries, and postings provide information about the job and its wage, above and beyond the skills requested.

Estimating wage premia for various job characteristics by applying hedonic regression has been common in the economics literature (Mincer, 1974; Heckman et al., 2006; Weinberger, 2014; Deming, 2017). Hedonic regression techniques uncover the predictive value of characteristics for equilibrium outcomes in the market. Because both sides of the market are heterogenous, the equilibrium prices provide information to both firms and workers.

This work builds on Autor and Handel (2013), Deming and Kahn (2018) and Marinescu and Wolthoff (2020), three pioneering papers that highlighted the wage heterogeneity within occupation and demonstrated that additional characteristics like tasks, skills demanded, and job titles can explain this variation.

Autor and Handel (2013) conduct a survey to collect new data on the job activities of a representative sample of U.S. workers across task domains, and demonstrate that within-occupation measures have significant and economically meaningful predictive power for earnings. This process relies on nationally representative survey data for a sample of 1,333 workers. The drawback of this approach is its lack of scale: to identify rare characteristics, the sample must be substantial. To identify differences over time, the survey must be conducted repeatedly.

Papers that followed used data from online job boards. The advantage of this approach is that these analyses can be done in closer to real-time, and avoid costly surveys. Deming and Kahn (2018) show that skill requirements affect average wages of professionals across MSAs, explaining up to 94% of the variation in average wages in MSA-occupation cells. The analysis focuses on average wages, when there is substantial variation *within* occupation in wages. Furthermore, the sample is understandably limited to professional job advertisements, as during that time period (2010-2015), online job postings leaned heavily towards professional occupations. Marinescu and Wolthoff (2020) find a coefficient of determination (R^2) of almost 90%, looking at the explanatory power of job titles using posted wages on Career Builder. This number is remarkably high, but is limited to the sample of under 20% of postings that posted wages. Given that postings with and without wages systematically differ, this may be difficult to extrapolate to the general population. My work extends this research by introducing a new dataset with salaries derived from the metadata of job postings. I also demonstrate that job titles have little *out-of-sample* predictive power because the number of unique job titles is very high.

Differing from previous interpretable approaches like high-dimensional fixed effect regressions, the natural language processing methods used in this paper often lack interpretability. In the context of this research, this implies that though the model can explain what differentiates a high and low salary job posting, it is difficult to translate this information into actionable insights. However, generating a counterfactual posting with the characteristic in question and treating the machine learning model's weights as a vector of coefficients can create the circumstances to interpret the implicit labor market value of the characteristic. This approach, first introduced in Bana, Brynjolfsson, Rock and Steffen (2021), injects additional text into thousands of postings and runs them through the model to recover an estimate of the valuation associated with the injected text.

I apply this text injection method to the question of online learning. With the rise of massive online open courses (MOOCs) and online certification programs, there are many opportunities to make small-scale, career relevant human capital investments. Furthermore, these investments may be more accessible to even larger swaths of the population. By nature, these programs are shorter, more narrowly focused, and offered in a flexible time frame, providing an avenue for workers who might be more constrained to invest in their own upskilling or retraining. While many of these endeavors may increase one's productivity, there is currently no clear way to understand the potential earnings consequences from each of these upskilling initiatives. Traditionally, large human capital investments have been rigorously evaluated using administrative data and randomized control trials (Athey et al., 2019; Altonji and Zhong, 2021). However, micro-credentials such as MOOCs and online certifications have proliferated at an unprecedented pace and volume.[1] If this pace is characteristic of the new era described in popular discussions as the "future of work," then the arrival of new skills and certifications may eclipse the ability to evaluate them through traditional mechanisms.

There is no centralized database of online certification programs and their costs. I take a sample of certifications from Indeed.com's "10 In-Demand Career Certifications." These certifications cost between $225 and $2050, not including time necessary to prepare. I find that these certifications carry a wide range of model predicted returns. While most of these certifications are associated with a significant and positive predicted effects on salaries, in some cases, these benefits may take more than one year to accrue, and may not even yield positive salary outcomes for some postings. For example, for the Cisco Certified Internetwork Expert (CCIE), one of the most prestigious networking certifications in the industry, the mean salary premium is 0.013 log points, conferring a mean benefit of less than $1000 in a single year, when the cost of the certification is over $2050 in fees alone. Furthermore, for over 25 percent of postings, the model predicts that the certification confers no positive salary benefit. On the other hand, the "IIBA Agile Analysis Certification" is associated with a 0.047 log point increase in salary, or $3140 at the mean of the salary distribution, implying the benefits exceed the costs in just a few months.

To the best of my knowledge, this research serves as the first independent estimates of the value of these certifications. While the professional associations and firms that authorize these certifications often advertise their value, they may suffer from an incentive compatibility problem. Moreover, my approach scales, potentially to the universe of skills and certifications, while also allowing for temporal, spatial, and occupational variation in the premia. This information can serve to further personalize recommendations to employers and job seekers, with major implications for information systems.

[1] Certification Magazine reports over 900 IT certifications in its 2020 survey.

2 Model

2.1 Model Structure

To describe the process by which the NLP model takes text as an input to predict salaries, it is helpful to think about the limits of traditional data for this analysis. Suppose our objective was to compare a group of job postings. We might transform this into traditional data by counting all the distinct words in each posting. The resulting matrix would be full of zeroes, as many postings would not contain certain words, creating challenges for traditional regression analysis. Moreover, the number of prepositions or conjunctions in each posting might not necessarily be meaningful. Even if the data was not sparse, simply counting words might be suboptimal: we improve the situation by counting pairs of words (called bigrams), instead of counting individual words because "learning machine" and "machine learning" have different implications. This logic might extend to trigrams or other n-grams. However, words that have differences in meaning when utilized in different contexts would be obscured through this method. For example, the word python could represent a programming language, or a reptile.

The computer science community has identified a solution to these problems through a language model called BERT. BERT stands for Bidirectional Encoder Representations from Transformers. In 2018, when released, Devlin et al. (2018) achieved state-of-the-art performance on a number of NLP understanding tasks. Briefly, BERT embeddings are trained on the entirety of English language Wikipedia and 500 samples of text called the Book Corpus. The model is trained using two tasks: (1) masked language modeling, where 15 percent of tokens are masked and BERT was trained to predict them using the context, and (2) next sentence prediction, where BERT was asked to predict if a particular next sentence was probable given the first sentence. The purpose of these tasks is to output vectors for tokens that rely on context.

That is, when ingesting a job posting, each word (or subword) will be given a 768 dimensional vector based on the words around it. We could imagine that based on context, the vector for the word python when used as a programming language might be near other programming languages or words about debugging code, while the vector for the word python when used to describe reptiles might be near words for other snakes, like "boa," or words like "grass" or "slither."

In this paper, I currently utilize "pre-trained" BERT embeddings. That is, the vectors that are applied to each token (word or part of word) are based on English language Wikipedia and the Book Corpus. There are many reasons why the embeddings from these sources can convey similar meaning: both job postings and Wikipedia are written for relatively general audiences. Unlike patents or highly technical documents where the words represented may not even exist on Wikipedia, most words from job postings – which describe responsibilities, firm attributes, team composition, and educational requirements – represent similar concepts as they would on Wikipedia. There may be cases where words in job postings convey different meanings than on Wikipedia or in unpublished books. For example, the terms "preferred" and "desired" may convey similar meanings in job postings but different meanings in romance novels. However, for most words in job postings, I would conjecture that the meaning would be similar on Wikipedia and in novels. This is a testable prediction, and future iterations of this work will develop pre-trained embeddings.

These pretrained BERT embeddings are the fundamental input of the NLP model predicting salaries, and therefore serve as the first layer. Because the BERT model has a length limit of 512 tokens, I only select the first 512 tokens of a job posting.[2]

More specifically, the model takes a job posting of 512 tokens as an input. A token in the BERT model is either a word, or a subword (part of the word), if the word is not sufficiently common. One estimate from another transformer model, GPT-3, suggests that, on average, 75 words consist of approximately 100 tokens. These 512 tokens are turned into a 512 by 768 dimensional matrix. This matrix is quite large, and the next layers in the model serve to reduce dimensionality. The model structure is displayed visually in Figure 4 and numerically in Table 1. First, a convolutional neural network summarizes each posting, by turning a single posting from 512 x 768 dimensions to 509 x 64 dimensions (taking four tokens at a time, conceptually

[2]In practice, this decision is not consequential: starting at a random point in the posting compared to starting at the beginning of the posting yields similar results. Longer postings seem to have more information about the application process and not actually about the job itself.

condensing separate words into phrases). The next layer is a global max pooling layer, which takes the maximum value over dimensions, resulting in a 64 dimensional vector per posting. The next two layers flatten and normalize, concluding with an output layer that predicts the salary. Greater discussion on the layers of the model and the hyperparameters are described in the Appendix.

2.2 Model Evaluation

The model is currently trained on 855,477 postings from April 2019 to December 2019. The relevant evaluation metrics are based on the 214,281 postings that are "out-of-sample," i.e. not used in the training process. In data science terminology, this can be referred to as the "test" sample.

Table 2 compares models that do not use the full text of data to the fifth model, described above, that uses the full text of data.

The first column is a model with six-digit occupation fixed effects provided by BGT. [3] The coefficient of variation (R^2) on a simple regression containing occupation fixed effects for the sample is 0.590. This is notably much higher than an individual or household level regression on earnings (instead of at the posting level). However, there is still much left to be explained.

The next model, in Column (2), incorporates location. BGT postings are tagged with a best fit metropolitan statistical area (MSA). A fully interactive model, with separate fixed effects for each occupation by MSA, would capture the variation discussed in Deming and Kahn (2018), allowing for different local labor markets to have different skill requirements (and therefore, different wages) for different occupations. This model yields an R^2 of 69.5 percent, around 10 percentage points higher than a model with only occupation fixed effects.

Previous work has suggested that the skills articulated in job postings have predictive power for wages. A number of papers, including but not limited to Acemoglu et al. (2020) and Deming and Kahn (2018), utilize the skill data from BGT to characterize differences within occupation across firm or MSA.

[3] Of the postings in the test sample, almost 95% of them are tagged with a six digit occupation label. The postings missing such a label are categorized as a separate category for the purpose of this analysis.

Along these lines, Column (3) estimates a random forest regressor with the 28 BGT skill cluster families, listed in Appendix Table A1. A random forest regression is an ensemble model, where a number of decision trees are averaged to make a more accurate prediction. A longer discussion of random forests will be in the appendix. In this circumstances, 1000 different decision trees were fit on the training sample, and evaluated on the test sample.

The resulting increase in the explanatory power of salaries is modest – from 0.695 to 0.728. However, the skill cluster families provide meaningful information.

Column (4) uses the more granular Burning Glass Technologies skill clusters. There are 648 different skill clusters. For example, under the skill cluster family Information Technology, there are skill clusters of "Cybersecurity," "Technical Support," and "Java." Under the skill cluster families of Maintenance, Repair, and Installation, there are the skill clusters of "Vehicle Repair and Maintenance," "Hand Tools," and "Electrical and Mechanical Labor." The coefficient of variation continues to increase – to 0.765, a 3.7 percentage point increase.

Finally, Column (5) describes the natural language processing model, outlined in detail in Section 2. The model performs substantially better – a coefficient of determination of 0.874. That is, the combination of words articulated in the job posting can explain 87 percent of the variation in metadata salaries, above and beyond the skills and occupation that come from the words. Quantifying this in another way, adding context (from Columns (4) to (5)) explains as much variation as adding geography (from Columns (1) to (2)).

The next row displays the same exercise for the Root Mean Square Error (RMSE). First, the RMSE falls only slightly between models (1) and (2), despite the sizable increase in the number of fixed effects. Moreover, the contrast between the fixed effects models and the NLP model is quite stark. The NLP model leads to a 42.9 percent decrease in the RMSE, compared to the occupation by MSA fixed effects specification.

3 Data

The data comes from two distinct data vendors, Greenwich.HR and Burning Glass Technologies. In this section, I describe the elements of the data used for each portion of the analysis.

3.1 Greenwich.HR Data

Greenwich.HR (GHR) is a labor market intelligence firm that provides real-time labor market data to application developers, analysts and consultants. They consolidate job postings from millions of different sources. A major advantage of the GHR data is that they have collected pay data for over 70 percent of job postings collected in recent months. Though the exact method by which GHR collects this data is proprietary, I outline the approach in general terms to lend credence to the estimates.

While many postings do not contain information on wages, it is common practice for job posting platforms to solicit salary data from the recruiter posting the job. For example, in Figure A1 Panel A and B, it can be seen on one popular platform, Indeed, that recruiters are encouraged to fill in either the exact rate, the range, a starting salary, or a maximum salary. This screenshot is for illustrative purposes only, as the platforms and methods for integrating data used by GHR are proprietary. Panel A suggests that this incentivizes applicants. In Panel C, a similar screen is included for LinkedIn.

This information can be found on the applicant side when searching for postings. Visualized in Figure A1 Panel D, a postings' salary band can be inferred by whether it appears in the search results when changing the pay threshold. These images are intentionally taken from different platforms to demonstrate the ubiquity of this practice.

Key for the analysis, the postings' salary band is drawn from the metadata of the posting, as opposed to the characteristics of the postings itself. That is, GHR does not create a mechanical correlation between the posting language and the salary reported.

This pay data provides a major asset for analysis. However, like many new datasets, there are limitations. First, GHR did not collect the raw job text until 2020. Second, GHR sought to be a comprehensive source of the U.S. economy only beginning in March 2019. Prior to this time period, the focus was on public firms and certain sectors. The first limitation can be overcome by connecting postings between Burning Glass Technologies, which does collect the full text of the posting, and GHR. The second limitation precludes time series analyses on the changing wage premia over time going back. Because the COVID19 pandemic occurred in 2020, likely changing the premia associated with certain skills, this work focuses on cross-sectional variation in wages during the year 2019.

GHR contains 62,026,448 job postings for the period April 2019 to September 2020 (18 months). Of these postings, 37,113,670 contain posted salaries (59.8 percent). The posting distribution is displayed in Figure 1. As evidenced by the jagged lines in the density distribution, posted salaries do bunch at round numbers.

3.1.1 Comparison to CPS

To the best of my knowledge, there is no source of nationally representative posted salaries to compare GHR data to determine potential selection issues. The best alternative is comparing the salary distribution to the distribution of weekly earnings in the Current Population Survey. The Current Population Survey (CPS) collects earnings from one-fourth of the monthly sample, limited to wage and salary workers. The closest comparison is usual weekly earnings, representing data before taxes and other deductions, and including any overtime pay, commission or tips usually received.

I use the fourth quarter in 2019's CPS release for this comparison, graphically depicted in Figure 2. The 25th percentile of CPS weekly earnings is $623, which at 52 weeks a year is $32396. This is quite close to the 25th percentile of GHR salaries, at $32175.19. The median CPS weekly value is $936, which is an annual value of $48,672. This is much lower than the GHR median of $41,750. This pattern continues, with the 75th percentile of CPS earnings is $77376 annually, while the GHR percentile is $66501.

There can be several reasons to expect the posting distribution and the actual salary distribution to differ. The two broad categories of reasons are (1) differences in job composition and (2) differences in reporting of pay.

The posting distribution represents new jobs, and therefore, industries and occupations that have higher turnover are likely to be overrepresented. For example, according to the BLS Job Openings and Labor Turnover Survey (JOLTS), the government sector has relatively low turnover, while the private sector has higher turnover. Within the private sector, there are also notable differences: leisure and hospitality is a high turnover industry, while durable good manufacturing is low turnover. Moreover, there are notable differences within occupations. In one extreme example, seasonal work has tremendous turnover, with large fractions of Lifeguards, ski patrol, and other recreational protective service workers being rehired at the beginning of every season. Given that higher turnover jobs are more likely to be lower wage, this is consistent with the overall directional difference between the posting distribution and the CPS distribution.

Differences in job composition between the posting distribution and the actual salary distribution can also be a function of how workers are hired. First, not all jobs are posted online. Previous research on online job postings has emphasized that as online job postings have become more common, firms and jobs added more recently are lower skilled (Blair and Deming, 2020). Moreover, not all jobs are posted and some postings may still represent more than one vacancy, despite the best attempts to deduplicate. To the best of my knowledge, there is no credible estimate of the fraction of jobs that are not posted, although ongoing work by researchers at the Bureau of Labor Statistics seeks to answer this question.

Though the job composition is likely different, the CPS and GHR are also measuring different underlying concepts. The CPS usual weekly wage includes expected overtime, commission and tips. These are not included in the GHR data.

The distributions are clearly different; however, it is difficult to assess whether this is a cause for concern. Future analyses will test robustness to various assumptions about the distribution.

3.1.2 Comparison between GHR Postings with and without Salaries

Another approach to assessing the representativeness of GHR salaries is to measure how much other observable characteristics can explain whether the salary exists. Using a 20 percent random subsample of postings from April 2019 to December 2019, I regress a binary for whether the salary is missing on six digit Standard Occupation Classification (SOC) code fixed effects. If occupations that are higher wage are less likely to be well represented with salary data, then occupation fixed effects should explain considerable variation in whether the salary is missing.

Instead, I find that the pseudo R^2 on a probit regression with occupation fixed effects is only 0.0134. That is, which postings have salaries in the data cannot be explained by the occupations of those postings. This is, by no means, conclusive evidence that salaries from metadata are random. However, it does suggest that the process by which salaries appear in metadata differs from what might be expected for posted salaries.

3.2 Burning Glass Technologies Data

Burning Glass Technologies (BGT) is an analytics software company that strives to provide real-time labor market information to higher education institutions, firms and municipalities. The product used in this analysis is the job postings data, collected from over 40,000 online job boards and company websites. These postings are deduplicated in a proprietary manner and the job title and employer name are cleaned.

For the analysis described, the key attribute of the data employed is the raw job text. This raw text has been seldom used in prior research, and contains virtually all the information that the applicant will see. The job text frequently contains information about the firm, the role, and the application procedure, though this is not systematic.

For illustrative purposes, the raw job posting text of two sample postings from October 2019 are displayed in Figure 3. Both postings use different terms to convey similar information. For example, in the first posting, responsibilities are outlined in the "Key Responsibilities" section, while these same thoughts are outlined in the second posting under the heading, "What would you do? The Specifics." Postings also differ in length, and some postings have some information about benefits and how to apply.

I link a GHR posting with a BGT posting using the firm name, job title, and date of posting. The two datasets are cleaned differently, so connecting them involves a fuzzy match. Typos and extraneous information are more likely to be at the end of the firm name or cleaned title, which means a string distance measure that weighs the beginning of the string is preferred. For this purpose, I use a Jaro-Winkler distance metric.

3.3 Certifications

To the best of my knowledge, there is no well-defined list of all career certifications.[4] For this reason, I use a variety of web sources to compile lists of certifications perceived as in demand or related to high salaries.

The primary set of analyses focus on Indeed.com's "10 In-Demand Career Certifications (And How To Achieve Them)," published in 2021. The advantage of this article is it includes estimated costs for certification exams. For example, the Project Management Professional (PMP) certification involves a fee of $405 to $555. These range substantially - from a few hundreds of dollars to the Cisco Certified Internetwork Expert certification requiring a $450 cost for a written exam, and $1,600 for a lab exam. A list of certifications from this article and their monetary costs are outlined in Table 3.[5] With this additional information, I test the hypothesis that the return to this certification exceeds this cost.

The certifications in Table 3 can be perceived as traditional: some of these certifications have existed for decades or longer.[6] A future set of analyses focus on newer certifications from CIO.com's "The 15 most valuable IT certifications today." These certifications, focused on topics such as cloud architecture, data visualization, and cybersecurity, are much newer on average.

These lists are not meant to be comprehensive - they are meant to be exemplars for further

[4]Certification Magazine conducts an annual salary survey, but these survey participants are limited to IT professionals, and therefore, only a fraction of the potential certifications available.

[5]Certifications from this article that do not explicitly list costs are not included.

[6]For example, the Cisco Certified Internetwork Expert lab exam was first administered in 1993.

applications of this approach.

4 Empirical Approach

4.1 Text Injection Experiments

The empirical approach leans heavily on Bana, Brynjolfsson, Rock and Steffen (2021), which describes a method called text injection to recover the relationship between the text and an outcome through an NLP model. The intuition is that after a model has been trained, the information from the model can be recovered in an interpretable way by adding text to the posting and seeing how it affects the predicted outcome.

Pedantically, the model trained above, with fixed weights, can be described as

$$Y = f(X|\beta)$$

where Y is the outcome, in this case the salary, X is the posting text, and β are the learned parameter vector of weights derived from the BERT layer and training from the process described in Section 2. Recall that β is high dimensional and contains many interaction terms, differentiating it from counting words.

Therefore, for a given posting i,

$$y_i = f(x_i|\beta)$$

Adding text to a posting, in this case, denoted as t_i, provides an additional input to the model. Therefore, the posting without added text can be described as

$$y_{i,0} = f(x_i, t_i = 0|\beta)$$

while the posting with added text is

$$y_{i,t} = f(x_i, t_i = t|\beta)$$

The outcome of interest is the average value of t on salary. This amounts to an expectation:

$$\mathbb{E}[f(x_i, t_i = t|\beta) - f(x_i, t_i = 0|\beta)]$$

By sampling from all postings a large number of times, these can be treated as independent and identically distributed (i.i.d) random variables, drawing on the Central Limit Theorem (CLT) for consistency and inference.

4.2 Discussion

The approach described above works only for marginal changes. If a posting drastically changes as a result of a text injection, the change cannot be interpreted as marginal, and therefore the CLT does not apply.

This concern is not only an econometric one, but also a practical one when thinking about the statements that can be evaluated using the text injection approach. For example, the approach would not be appropriate for occupational licenses. A license can be considered a mandatory certification. Specifically, it is a state issued credential that a worker *must* possess to legally work for pay (Friedman, 1962). For example, a physician without a physician's license cannot perform the vast majority of physician responsibilities. Occupational licenses, which are required for entry into particular occupations, are distinct from certifications described in this study–which represent *marginal* changes in responsibilities or capabilities.

4.3 Identifying Appropriate Counterfactuals

Adding a Certified Business Analysis Professional certification to a Light Truck Driver may not be appropriate because no jobs within this occupation require this certification. Though it may raise or reduce the value of the posting, these values are less pertinent to the worker or firm likely making a decision about the value of the certification. For this reason, I create two categories of counterfactuals. The first is broad and the second is more narrow: (i) All postings, (ii) postings in occupations that include this certification.

An occupation is considered in the category (ii) control group if, at any time in the first quarter of 2019, there was a posting that BGT tagged as including this certification. The time period identified is intentionally distinct from the period of analysis to ensure that there is no mechanical correlation between the postings identified as requesting these certifications to create the control group and the analysis sample.

It is important for interpretation purposes to remember that an occupation is in the control group if a certification was mentioned in any of the occupation's postings. This does not mean that this was a requirement for the job. No current work using large scale data has distinguished between characteristics (such as certifications or skills) placed in a "Desired" and "Required" section of a posting. This is because while these distinctions appear in some job postings, they do not appear in all job postings.[7]

Some certifications are much more common than others in job postings. For example, "Project Management Professional (PMP)" is connected to 32,745 job postings in 254 distinct occupations in the first quarter of 2019. On the other hand, "Certified in Logistics, Transportation and Distribution" is connected to 75 job postings with 25 distinct occupations. The precision of the estimates will reflect these differences.

This category is created based on binaries - whether an occupation at any point had a request for this certification. This may place too much weight on false positives. An alternative approach, results forthcoming, adjusts this threshold based on the fraction of postings in this occupation that sometimes request this certification. This continues to reduce the sample size, but may represent a more representative counterfactual.

This approach also naturally lends itself to identifying heterogeneous treatment effects. Fu-

[7]Schema.org, a collaborative, community activity, which creates, maintains, and promotes schema for structured data on the internet does not separate out fields in the job posting schema for desired and required skills. As this is one of the major efforts to promote structure, it is unlikely that this distinction will be made in multi-platform job text analysis in the near future.

ture analyses, with a larger sample, will examine the effect of adding certifications to each occupation.

The advantage of this text injection approach, compared to just adding indicators for BGT skills or certifications to a regression on GHR salaries, is that the dimensionality reduction is done by the natural language processing model. Previous papers, such as Deming and Kahn (2018), explicitly categorize groups of skills out of the tens of thousands of skills that BGT tags data with. Another alternative is performing some sort of LASSO regression. This approach also performs dimension reduction but with far less context.

4.4 Evaluating over a time horizon

A certification is an asset that carries over beyond a single year. While some certifications require regular renewal (IIBA-AAC requries renewal every three years), some allow you to carry the designation for life. I estimate the value of the certification over three time horizons: one year, five years, and ten years. This amounts to a net present value (NPV) calculation of

$$NPV = \sum_{t=1}^{n} \frac{R_t}{(1+i)^t}$$

for $n = 1, 5,$ or 10. The discount rate for human capital is variable over time, and likely beyond the scope of this paper.[8]

5 Results

I begin with a common certification, the Project Management Professional (PMP) certification. This is considered the world's leading project management certification. It is administered by the Project Management Institute (PMI), and their website suggests that the median salary for U.S. project professionals is 25% higher with the PMP certification.[9]. The empirical results tell a substantially different story, displayed visually in Figure 5 and numerically in Table 4. In

[8]For a fascinating example where the value of a skill depreciates precipitously, see Horton and Tambe (2020).

[9]https://www.pmi.org/certifications/project-management-pmp/earn-the-pmp/why-the-pmp/pmp-earning-power

Figure 5 Panel A, the full sample of control postings (40,631 randomly selected postings) have Project Management Professional (PMP) added. Though this appears to make a slight difference, the difference is not large (the model predicts a salary increase at the median of 0.007 log points higher.) Though the increase is slightly larger for postings at the lower end of the salary distribution, the slope is only -0.205. The exercise is repeated with only occupations that have the certification requested in the first quarter of 2019. This control group is has a slightly higher salary on average, excluding some of the postings on the lower end of the salary distribution. On the other hand, this is still a large majority of the sample (77.7% of the postings fall into one of the occupations that sometimes requires a PMP certification). This means it is unsurprising that the results don't differ substantially across Panels A and B. The salary increase estimated in this table amounts to an average increase of 0.012 log points – for a posting at the mean, this amounts to a $580 increase in salaries. Given the exam cost, the certification benefits may not exceed the cost for a non-trivial fraction of the sample for the first year.

The "Certified Associate in Project Management (CPAM)" is another project management certification, geared towards entry-level workers. This certification displays the largest increase in salaries within the full sample. Unfortunately, BGT does not collect data on the CAPM. For this certification, I search for this expression in the full text of postings collected by BGT. For the full sample, this certification increases the posted salary by a substantial amount, 0.073 log points. The CAPM is a junior level certification, granted by the same institute that grants the PMP. Thus, we would expect the increase from the addition of the CAPM to be greater at the lower end of the distribution, at least compared to the PMP. Indeed, the difference is highly negatively correlated between the difference and the original, with a correlation coefficient of -0.477.

However, the sample of postings requesting the CAPM is on average higher salary than the sample of postings only requesting the PMP. This is not expected, and may be associated with the term Project Management Professional not necessarily denoting the certification in a posting. Far fewer occupations are associated with the CAPM. Adding the CAPM brings about a significant and positive salary increase. This is a very lucrative credential, with the upper bound on the cost being $300, and the salary increase at the mean of $3401.54.

The following two certifications, "Certified Business Analysis Professional (CBAP)" and "IIBA Agile Analysis Certification (IIBA - AAC)" are business analyst certifications. In fact, they are both granted by the same professional organization, the International Institute of Business Analysis. These certifications differ in that the first one, the CBAP, is geared towards more senior professionals, whereas the AAC is geared toward "Agile" methods. The mean effect of the CBAP is 0.025 log points, with a small increase even at the 25th percentile of the distribution. The control group of only occupations that have asked for a CBAP designation is much higher wage. The mean increase between the full sample of postings and the smaller control group is almost 0.01 log points, implying the naive comparison may be an overestimate. Even then, the 0.016 log point increase of the CBAP is a statistically significant increase in salary (s.e. 0.00025).

Consistent with the seniority levels implied by the certifications, the average salary posted for the CBAP is higher than the average salary posted for the IIBA-AAC, though these differences are not statistically significant. However, the IIBA seems to have a much larger effect on salaries. The mean salary increase in the full sample is 0.06 log points, while the mean salary increase in selected occupations is 0.047, still a large increase. On the sample of workers at the mean log salary, this amounts to a $3140 increase, a substantial return on investment taking into account the cost of the exam, even in the first year. The percentiles of the distribution suggest that even the 25th percentile of the effect is positive, at 0.015 log points.

Given the prevalence of agile methodologies in the past quarter century (Rigby et al., 2016), the term agile itself may itself have consequences for a job posting. Further work with integrated gradients methods may be able to test this hypothesis.

The third category of certifications is associated with supply chain. Given the growth of the warehousing and courier sectors described in Choe et al. (2020), and the rise in e-commerce, this is a skill set in the economy receiving a lot of attention. Like the Business Analysis certifications, these three are available through the same professional association, the Association for Supply Chain Management. The first is "Certified in Production and Inventory Management

(CPIM)," geared towards increasing an organization's profitability and optimizing production and inventory management within the organization. The inclusion of this designation increases salaries by 0.013 log points (s.e. 0.0002) in the full sample of postings, and 0.008 log points (s.e. 0.0002) in the relevant occupation sample. Despite these being relatively small effects, they continue to be statistically significant.

The Certified Supply Chain Professional (CSCP) certification requires a bachelor's degree, three years of experience, or the CPIM or one of many other certifications as a prerequisite. It is reassuring to see, therefore, that the average salary from the CSCP sample is higher than the CPIM sample. Similar to the CPIM, the CSCP has a statistically significant effect on earnings, though this effect is small. At the mean, the increase is approximately $300, suggesting it takes more than one year to receive a positive return on investment. This is a stark contrast to the IIBA-AAC, which pays for itself in months. The third supply chain certification "Certified in Logistics, Transportation and Distribution (CLTD)," is focused on warehouse and transportation fundamentals. Similar to the CPIM, it does not require a bachelors degree. However, the salary distribution is much higher than the other two control groups. The results from the CLTD are similar to the above two certifications. Though the certifications add value, as evidenced by the significantly higher mean salary, this value comes much closer to the cost of the certification.

The final category of certifications are computer network, both by Cisco. The first one, "Cisco Certified Internetwork Expert (CCIE)," requires a written exam combined with an eight hour lab exam. According to one training provider, this is perceived as the toughest certification to achieve. Table 11 demonstrates that on average, however, this is not as lucrative as it appears. Based on postings, mentioning the CCIE increases salaries by 0.013 log points, with over 25 percent of postings experiencing no salary improvement with the mention. The "Cisco Certified Network Professional" is a lesser version of the CCIE. Once again, it is reassuring that the average salary of the occupations that request the CCIE are higher those that request the CCNP. The CCNP is associated with a quite similar return, 0.012. Both the CCIE and CCNP have tracks,and it is possible to look individually at these different tracks to see whether there are some that are more valuable than others. In either case, these estimates of the mean salary increase are statistically significant and positive. For the CCNP, at the mean, the salary is estimated to increase by $733, an amount that exceeds the upper bound fee.

These results are summarized in Figure 7. Both the benefits and the costs of these certifications vary, and these are uncorrelated both for predicted change in log salary, and in dollars.

The nine certifications described in this analysis have all been considered "In Demand" by a popular job search website. Yet the premia associated with each of these certifications are significantly different. Moreover, estimates range from 0.005 log points to 0.048 log points – almost a tenfold difference. The real-time pricing of these attributes can provide additional information to firms and workers about how to strategically invest, improving decisions about human capital accumulation.

Though thus far, the examination has been on certifications that are well recognized, this approach extends to new certifications, skills, and other marginal attributes. With the rise of learning opportunities, this method provides an approach for information at scale.

6 Conclusion

This paper develops the first natural language processing model to predict the salary of job postings using the text. With new data on salaries from the metadata of job postings, the inputs and outputs are well-defined. This lends itself to the task of supervised machine learning, where the task is to derive the function that relates text to salaries. Because text in job postings is written in commonplace language, I use the technique called transfer learning – applying knowledge gained from solving one problem to apply to this problem of salary prediction. In practice, this means that the first layer of my salary prediction model is pre-trained word embeddings from the BERT model, trained on English language Wikipedia and the Book Corpus.

My model substantially exceeds performance by any conventional baseline – a 43 percent decrease in RMSE and a 26 percent increase in R^2 compared to models with occupation by MSA fixed effects. This demonstrates that variation important for earnings can be found in the text of online job postings.

To apply the exceptional performance of the model to critical questions about job attributes like skills and amenities, I employ an approach of developing counterfactual postings. The difference between the predicted salary of the counterfactual posting (with the marginal characteristic) and the original posting's predicted salary yields the price of the marginal characteristic. This price can differ across postings so the appropriate counterfactual distribution can be defined depending on the context of the problem at hand.

This approach lends itself to a myriad of research questions. I demonstrate the application to "In-Demand" certifications, an important question related to upskilling. The methods described can also answer questions about changes in the prices of information technology skills, job amenities, and with appropriately identifying variation, the causal effect of policies on salaries.

References

Acemoglu, D., D. Autor, J. Hazell, and P. Restrepo (2020): "AI and Jobs: Evidence from Online Vacancies," Working Paper 28257, National Bureau of Economic Research.

Altonji, J. G. and L. Zhong (2021): "The Labor Market Returns to Advanced Degrees," *Journal of Labor Economics*, 39, 303–360.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019): "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," Working Paper 26463, National Bureau of Economic Research.

Autor, D. H. and M. J. Handel (2013): "Putting Tasks to the Test: Human Capital, Job Tasks, and Wages," *Journal of Labor Economics*, 31, S59–S96.

Blair, P. Q. and D. J. Deming (2020): "Structural Increases in Demand for Skill after the Great Recession," *AEA Papers and Proceedings*, 110, 362–65.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. (2021): "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*.

Choe, D., A. Oettl, and R. Seamans (2020): "What's Driving Entrepreneurship and Innovation in the Transport Sector?" Working Paper 27284, National Bureau of Economic Research.

Deming, D. and L. B. Kahn (2018): "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals," *Journal of Labor Economics*, 36, S337–S369.

Deming, D. J. (2017): "The Growing Importance of Social Skills in the Labor Market*," *The Quarterly Journal of Economics*, 132, 1593–1640.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

Domo (2018): "Data never sleeps 6: Domo," .

Friedman, M. (1962): "Occupational licensure," *Capitalism and freedom*, 137.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006): "Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," Elsevier, vol. 1 of *Handbook of the Economics of Education*, 307–458.

Horton, J. J. and P. Tambe (2020): "The death of a technical skill," *Unpublished Manuscript*.

Kuhn, P. and H. Mansour (2014): "Is Internet Job Search Still Ineffective?" *The Economic Journal*, 124, 1213–1233.

Marinescu, I. and R. Wolthoff (2020): "Opening the Black Box of the Matching Function: The Power of Words," *Journal of Labor Economics*, 38, 535–568.

Mincer, J. (1974): *Schooling, experience, and earnings*, National Bureau of Economic Research.

Rigby, D. K., J. Sutherland, and H. Takeuchi (2016): "Embracing agile," *Harvard business review*, 94, 40–50.

Weinberger, C. J. (2014): "The Increasing Complementarity between Cognitive and Social Skills," *The Review of Economics and Statistics*, 96, 849–861.

Figures and Tables

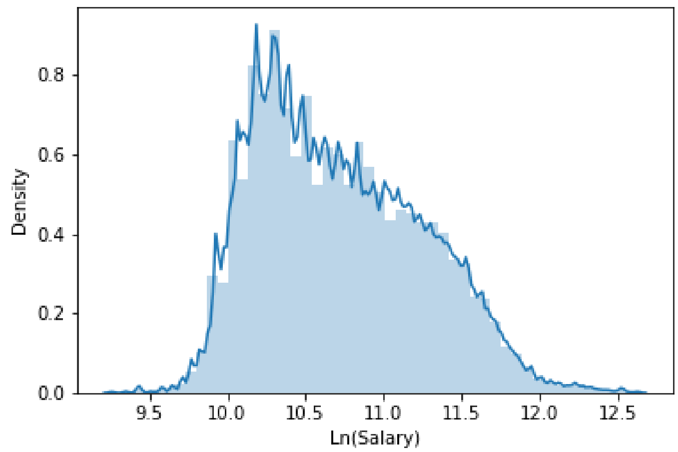


Figure 1: GHR Salary Distribution from April 2019 to September 2020

Notes: This figure describes the posted salary distribution of the 37,113,666 Greenwich.HR job postings with salary metadata posted between April 2019 and September 2020. The mean of the distribution is 52473.28.

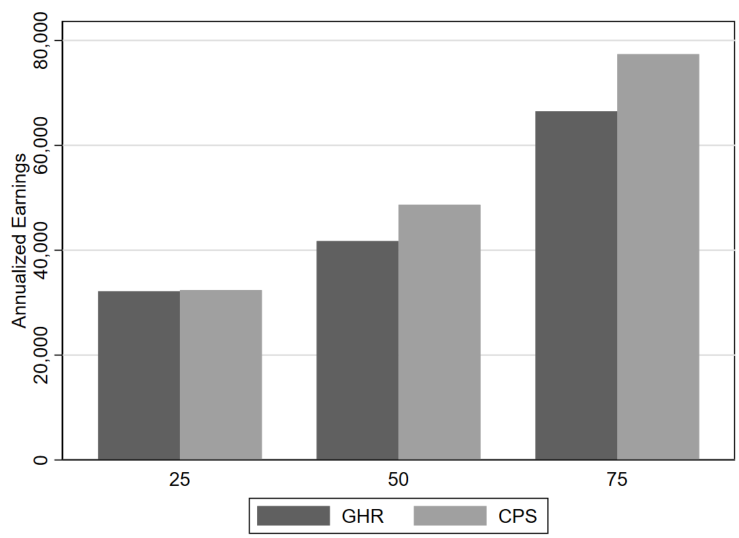


Figure 2: CPS - GHR Comparison

Notes: The Current Population Survey (CPS) values for quartiles of weekly earnings come from the Bureau of Labor Statistics' Usual Weekly Earnings of Wage and Salary Workers News Release Fourth Quarter 2019, available at `https://www.bls.gov/news.release/archives/wkyeng_01172020.htm`. CPS earnings are annualized by multiplying by 52. Data represent earnings before taxes and other deductions and include any overtime pay, commissions, or tips usually received (at the main job in the case of multiple jobholders). Greenwich.HR (GHR) salaries come from the full set of 37 million postings with salary metadata available.

Figure 3: Sample Job Postings in the Portland - Vancouver - Hillsboro Metropolitan Statistical Area

(a) Sales Floor Associate at Buy Buy Baby

```
'Sales Floor Associate\n\nBuy Buy BABY\n\n-\n\nBeaverton, OR 97005
\n\nThe Sales Floor Associate oversees a Department within the sto
re. In this role you will be a product, service and selling expert
for your area while meeting sales and productivity goals.\nKey Res
ponsibilities:\n\n* Models exceptional customer service by buildin
g relationships with store customers; makes appropriate recommenda
tions based on customer needs; drives sales through suggestive sel
ling, add-ons, and home deliveries\n* Meets with customers on a on
e-on-one basis to assist with determining personal needs and compi
ling merchandise preference list\n* Explains features of a broad a
rray of merchandise to customers\n* Promptly and politely responds
to customer inquiries and requests for support\n* Resolves custome
r issues using customer service skills, and escalates issues to mo
re senior associates as necessary to ensure customer satisfaction
\n* Organizes and straightens merchandise areas on the sales floor
\n* Performs Registry Specialist tasks\n* Performs Sales Associate
tasks\n* Knowledgeable of available technology and tools\n* Assist
s customers by offering a Baby order when merchandise is out of st
ock or not carried in the store\n* Performs additional duties as r
equired including, but not limited to, stocking, freight processin
g, price changes, cart retrieval, break room and restroom housekee
ping\n* Demonstrates commitment to the organization by maintaining
regular, on site attendance, is reliable and follows through with
responsibilities\n\nEducation/Experience:\n\n* High School diploma
or equivalent\n* 2-4 years of retail experience desired\n\nsave th
is job a'
```

(b) Sales Associate at National Vision Inc.

```
'Sales Associate\n\nNational Vision, Inc.\n\n-\n\nVancouver, WA 98
684\n\nPosition Description:\n\nAt National Vision, we believe eve
ryone deserves to see their best to live their best. We help peopl
e by making quality eye care and eyewear more affordable and acces
sible.\n\nNational Vision, Inc. (NVI) is one of the largest optica
l retailers in the United States. We offer an innovative culture w
here training is a priority, hard work is praised, and career grow
th is a reality.\n\nWe are looking for a Sales Associate to join o
ur growing team. The Sales Associate is responsible for selling, f
itting and dispensing eyewear to customers.\n\nWhat would you do?
The Specifics\n\n* Meet NVIs sales and company objectives.\n* Foll
ow the Americas Best Code of Excellence to ensure customer satisfa
ction by creating a warm and welcoming environment for customer
s.\n* Assist with dispensing eyeglasses and contact lenses to cust
omers, as permitted by state law.\n* Perform insertion and removal
training of contact lenses to customers, as permitted by state la
w.\n* Educate clients on proper eyeglass and contact lens care.\n*
Maintain accurate and organized patient records.\n* Assist Optomet
ric Technician, Receptionist, and Contact Lens Technician when nec
essary.\n* Answer, screen, and forward incoming phone calls in acc
ordance with NVI protocol.\n* Maintain visual merchandising accord
ing to Brand and Company Standards.\n\nPosition Requirements:\n\n*
Previous retail experience preferred, but not required.\n* Maintai
n license, as required by state.\n* Strong selling skills, aimed a
t meeting both the stores and self-sales targets, by following com
pany policies.\n* Strong customer service skills.\n* Able to give
instruction in a clear and concise manner to customers.\n* Effecti
ve interpersonal skills.\n* Excellent organizational skills.\n* De
tailed oriented.\n* Multitasking and time-management skills.\n* Ab
ility to learn optical knowledge.\n* Professional attitude and app
earance.\n* In some locations, bilingual abilities desired.\n\nWha
t are the benefits?\n\nNational Vision offers a competitive benefi
ts package including Health and Dental Insurance, 401k with compan
y match, Flex Spending Account, Short Term and Long Term Disabilit
y Insurance, Life Insurance, Paid Personal Time Off, and much mor
e. Please see our website at www.nationalvision.com to learn mor
e.\n\nsave this job a'
```

Notes: The job text of two sample postings in raw form from Burning Glass Technologies.

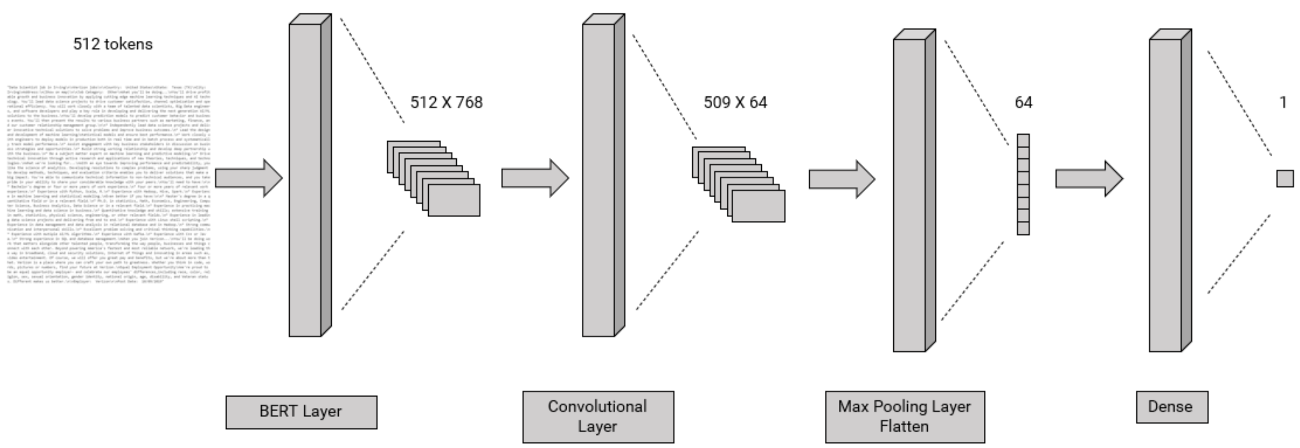
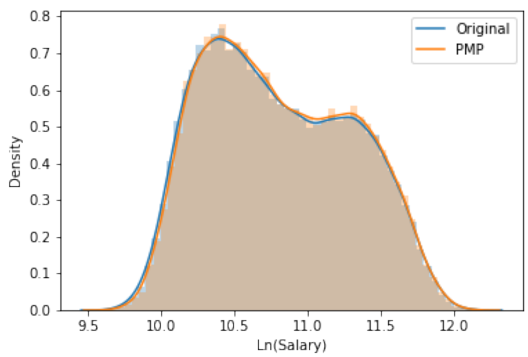


Figure 4: Model Structure

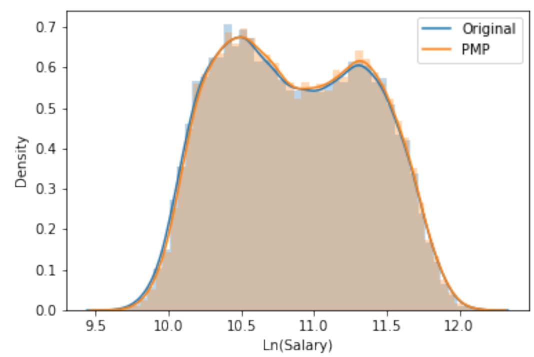
Notes: This figure displays the model structure. A job posting of 512 tokens is the input. BERT embeddings take these 512 dimensions and assign each a 768 dimensional vector depending on context. The next layer is a one dimensional convolutional layer. It is ultimately identifying n-grams that are predictive. The resulting matrix is 509 x 64. The next layer is a global max pooling layer, which captures the most relevant features from a sentence. This layer is flattened and turned into a 64 dimensional vector, which eventually predicts one dimensional ln(salary). The parameters are also laid out numerically in Table 1.

Figure 5: Model Predictions for Certification: Project Management Professional (PMP)

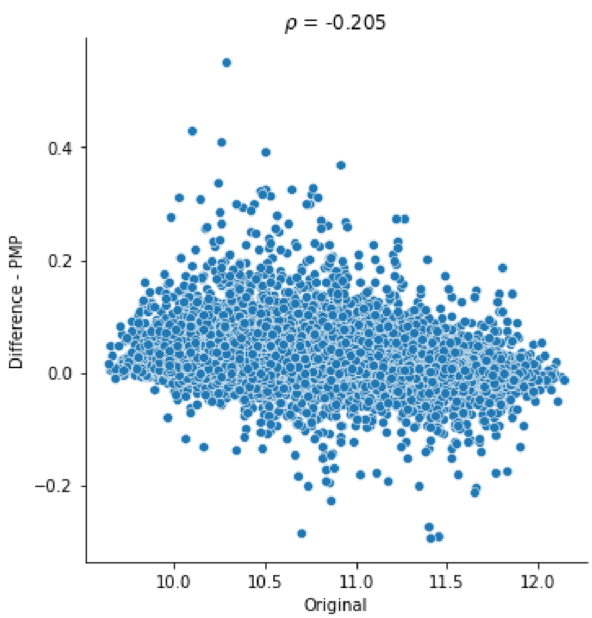
(a) Salary Distribution with and without Certification for Full Sample of Postings



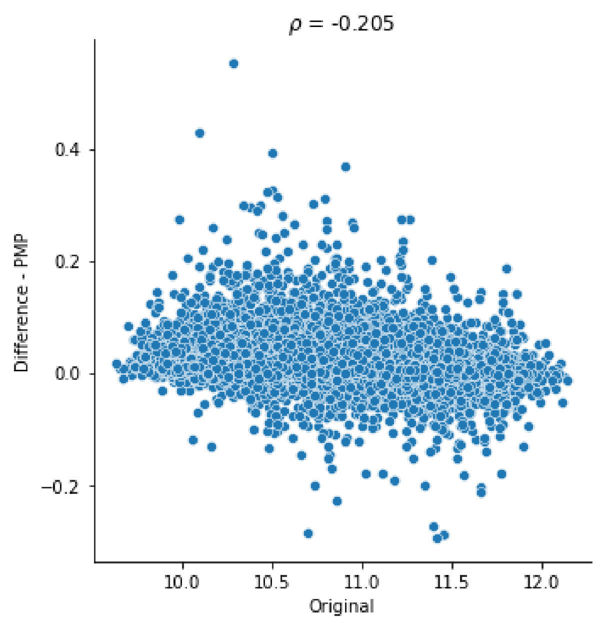
(b) Salary Distribution with and without Certification Limited To Occupations that Request Certification



(c) Correlation Between Original Salary and Difference for Full Sample



(d) Correlation Between Original Salary and Difference for Selected Occupations

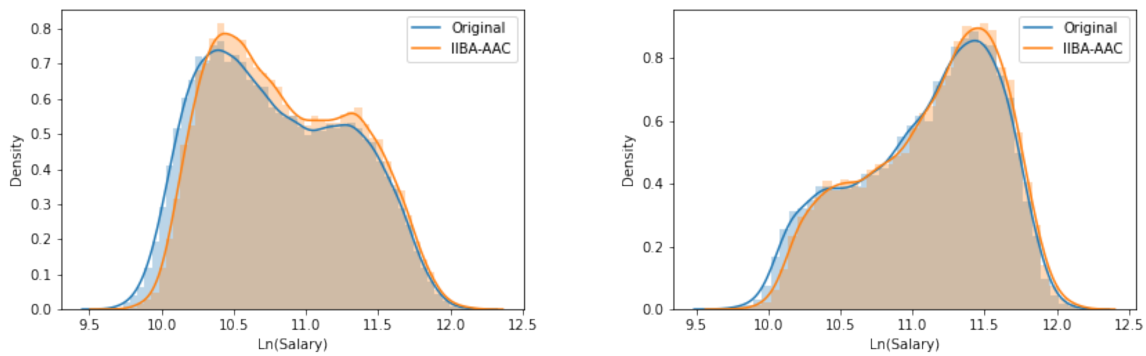


Notes: This figure demonstrates the model output and text injection results for the certification, "Project Management Professional (PMP)." Panel A shows the salary distribution with and without certification for the full sample of postings. Panel B shows the distribution, limited to occupations that at any point in the first quarter of 2019 ask for the PMP certification. Figures C and D demonstrate the relationship between the difference predicted by the model for the text injection and the original predicted salary. This difference is mildly negative for both the full sample, and the sample of occupations that have asked for certification. The correlation coefficient is also the same.

Figure 6: Model Predictions for Certification: IIBA Agile Analysis Certification

(a) Salary Distribution with and without Certification for Full Sample of Postings

(b) Salary Distribution with and without Certification Limited To Occupations that Request Certification

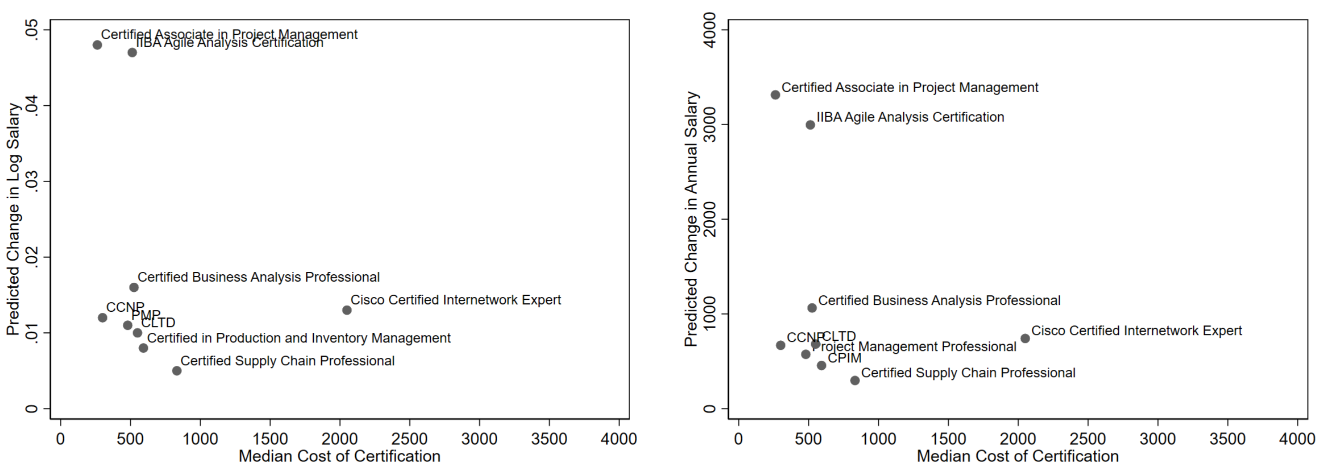


Notes: This figure demonstrates the model output and text injection results for the certification, "IIBA Agile Analysis Certification." Panel A shows the salary distribution with and without certification for the full sample of postings. Panel B shows the distribution, limited to occupations that at any point in the first quarter of 2019 ask for the IIBA certification.

Figure 7: Costs and Benefits

(a) Benefits in Log Points

(b) Benefits in Dollars



Notes: This table summarizes the relationship between cost of certification and benefit of certification, as measured by the model described in the draft. Median cost of certification is calculated as the median from Table 3. Benefits are in Tables 4 through 12. Dollar values are calculated at the mean of ln(salary) for only the occupations that ask for the credential (Panel B in those tables).

Table 1: Model Architecture

Layer Type	Dimensions	Number of Parameters
Input Layer	(X, 512)	
BERT Layer	(X, 512, 768)	109482240
Convolutional Layer	(X, 509, 64)	196672
Global Max Pooling Layer	(X, 64)	
Flatten	(X,64)	
Batch Normalization	(X, 64)	256
Dense Layer	(X, 64)	4160
Dense Layer	(X, 1)	65
Total params: 109,683,393		
Trainable params: 201,025		
Non-trainable params: 109,482,368		

Notes: This table describes the architecture of the natural language processing model used to predict salaries. In this table, X denotes the number of postings fed into the model. The input is 512 tokens of a job postings from October 2019. These 512 tokens are fed into a BERT embedding layer, where each token is given a 768 dimensional vector that is context dependent. At this point, each posting has 512 x 768 dimensions – likely too many inputs to a single salary value, so the next layers are focused on condensing dimensionality. The first step is a convolutional layer, which takes 512 x 768 dimensions, and reduces it to 509 x 64. The next layer, a global max pooling layer, takes the maximum values from this 509 x 64 matrix, which can be perceived as the most salient features, and condenses it to just 64 dimensions. The following two layers flatten and normalize these layers. Eventually, these 64 dimensions are condensed to a single dimension - the natural log of salary.

Table 2: Out-of-Sample Performance Metrics

	(1) Occupation FEs	(2) Occupation x MSA FEs	(3) Occupation, MSAs, & Skill Cluster Families	(4) Occupation, MSAs, & Skill Clusters	(5) NLP Model
R^2	0.590	0.695	0.728	0.765	0.874
RMSE	0.330	0.317	0.269	0.249	0.181
Occupations	785	785	785	785	
Locations	-	807	807	807	
Skill Categories			28	648	

Notes. This table summarizes the performance of the natural language processing model, in Column (5), to a number of relevant baselines. Relevant metrics are R^2 (coefficient of variation) and Root Mean Square Error (RMSE). The entire test set (214,281 observations) is used in every model. The outcome is ln(salary). Column (1) includes six digit Standard Occupation Classification (SOC) fixed effects. Column (2) interacts these occupation fixed effects with MSA fixed effects. Column (3) estimates a random forest regressor model with Burning Glass Technologies Skill Cluster Family categories. These are listed in the Appendix. Column (4) replaces the skill cluster family categories with Skill Clusters. There are 648 skill clusters. Column (5) is the model described extensively in Section 2.

Table 3: Certifications For Analysis

Category	Abbreviation	Certification Title	Cost (Lower)	Cost (Upper)
Project Management	PMP	Project Management Professional	$405	$555
Project Management	CAPM	Certified Associate in Project Management	$225	$300
Business Analyst	CBAP	Certified Business Analysis Professional	$475	$575
Business Analyst	IIBA-AAC	IIBA Agile Analysis Certification	$450	$575
Supply Chain	CPIM	Certified in Production and Inventory Management	$495	$690
Supply Chain	CSCP	Certified Supply Chain Professional	$695	$969
Supply Chain	CLTD	Certified in Logistics, Transportation and Distribution	$475	$625
Computer Network	CCIE	Cisco Certified Internetwork Expert	$2050	$2050
Computer Network	CCNP	Cisco Certified Network Professional	$300	$300

Notes: This list comes from the Indeed.com article, "10 In-Demand Career Certifications (And How To Achieve Them)," published by the Indeed Editorial Team on July 23, 2021. If costs are separated into application fees and other costs, the columns with cost reflect the total amount. This table only contains entries for which the cost of the certification has been included.

Table 4: Salary Predictions from the Addition of "Project Management Professional (PMP)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
With PMP	40631	10.806	0.497	9.647	10.392	10.751	11.216	12.129
Difference	40631	0.012	0.028	-0.294	-0.001	0.007	0.019	0.552
Panel B: Only Occupations That Ask for Credential								
Original	31550	10.867	0.503	9.632	10.445	10.845	11.292	12.143
With PMP	31550	10.878	0.499	9.647	10.460	10.856	11.297	12.129
Difference	31550	0.011	0.027	-0.294	-0.002	0.007	0.018	0.552

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Project Management Professional (PMP) to the posting. The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a Project Management Professional (PMP) certification in the first quarter of 2019 (254 different standard occupations).

Table 5: Salary Predictions from the Addition of "Certified Associate in Project Management (CAPM)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
CAPM	40631	10.866	0.470	9.739	10.472	10.819	11.256	12.125
Diff	40631	0.073	0.076	-0.285	0.017	0.055	0.111	0.755
Panel B: Only Occupations That Ask for Credential								
Original	12377	11.166	0.437	9.736	10.869	11.236	11.506	12.143
CAPM	12377	11.213	0.410	9.877	10.938	11.285	11.530	12.125
Diff	12377	0.048	0.066	-0.285	0.006	0.029	0.075	0.657

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Certified Associate in Project Management (CAPM). The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that include the text "Certified Associate in Project Management" in the first quarter of 2019 (46 different standard occupations).

Table 6: Salary Predictions from the Addition of "Certified Business Analysis Professional (CBAP)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
CBAP	40631	10.818	0.489	9.677	10.410	10.761	11.223	12.147
Diff	40631	0.025	0.038	-0.287	0.003	0.015	0.037	0.551
Panel B: Only Occupations That Ask for Credential								
Original	15401	11.112	0.439	9.781	10.787	11.175	11.453	12.143
CBAP	15401	11.127	0.431	9.834	10.805	11.190	11.463	12.147
Diff	15401	0.016	0.031	-0.287	0.000	0.010	0.025	0.315

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Certified Business Analysis Professional (CBAP). The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a Certified Business Analysis Professional certification in the first quarter of 2019.

Table 7: Salary Predictions from the Addition of "IIBA Agile Analysis Certification"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
IIBA-AAC	40631	10.854	0.481	9.679	10.450	10.798	11.252	12.190
Diff	40631	0.060	0.061	-0.287	0.021	0.045	0.085	0.954
Panel B: Only Occupations That Ask for Credential								
Original	13428	11.086	0.491	9.706	10.720	11.177	11.479	12.143
IIBA-AAC	13428	11.133	0.473	9.776	10.776	11.220	11.511	12.190
Diff	13428	0.047	0.055	-0.287	0.015	0.035	0.066	0.692

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding IIBA Agile Analysis Certification to the posting. The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a IIBA Agile Analysis Certification in the first quarter of 2019.

Table 8: Salary Predictions from the Addition of "Certified in Production and Inventory Management (CPIM)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
CPIM	40631	10.806	0.489	9.673	10.399	10.753	11.209	12.128
Diff	40631	0.013	0.036	-0.657	-0.003	0.010	0.025	0.469
Panel B: Only Occupations That Ask for Credential								
Original	21680	10.954	0.490	9.632	10.537	10.972	11.362	12.143
CPIM	21680	10.963	0.478	9.696	10.555	10.983	11.360	12.128
Diff	21680	0.008	0.036	-0.361	-0.006	0.007	0.022	0.469

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Certified in Production and Inventory Management (CPIM) to the posting. The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a Certified in Production and Inventory Management certification in the first quarter of 2019.

Table 9: Salary Predictions from the Addition of "Certified Supply Chain Professional (CSCP)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
CSCP	40631	10.809	0.495	9.654	10.398	10.755	11.218	12.140
Diff	40631	0.015	0.031	-0.292	0.001	0.010	0.025	0.562
Panel B: Only Occupations That Ask for Credential								
Original	19641	10.995	0.489	9.656	10.581	11.047	11.398	12.143
CSCP	19641	11.000	0.483	9.668	10.593	11.051	11.397	12.127
Diff	19641	0.005	0.027	-0.290	-0.004	0.005	0.016	0.559

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Certified Supply Chain Professional (CSCP) to the posting. The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a Certified Supply Chain Professional certification in the first quarter of 2019.

Table 10: Salary Predictions from the Addition of "Certified in Logistics, Transportation and Distribution (CLTD)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
CLTD	40631	10.811	0.489	9.680	10.402	10.759	11.214	12.134
Diff	40631	0.017	0.038	-0.351	-0.000	0.014	0.031	0.476
Panel B: Only Occupations That Ask for Credential								
Original	8908	11.135	0.449	9.821	10.792	11.192	11.500	12.143
CLTD	8908	11.145	0.437	9.847	10.811	11.199	11.502	12.134
Diff	8908	0.010	0.036	-0.296	-0.006	0.009	0.025	0.382

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Certified in Logistics, Transportation and Distribution (CLTD) to the posting. The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a Certified in Logistics, Transportation and Distribution certification in the first quarter of 2019.

Table 11: Salary Predictions from the Addition of "Cisco Certified Internetwork Expert (CCIE)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
CCIE	40631	10.809	0.495	9.654	10.398	10.755	11.218	12.140
Diff	40631	0.015	0.031	-0.292	0.001	0.010	0.025	0.562
Panel B: Only Occupations That Ask for Credential								
Original	18479	10.957	0.514	9.632	10.517	10.974	11.401	12.143
CCIE	18479	10.970	0.507	9.654	10.534	10.989	11.407	12.140
Difference	18479	0.013	0.030	-0.292	-0.001	0.009	0.022	0.562

Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Cisco Certified Internetwork Expert (CCIE) to the posting. The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a Cisco Certified Internetwork Expert certification in the first quarter of 2019.

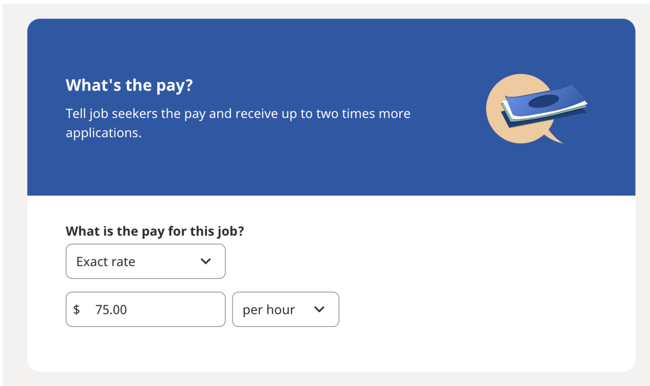
Table 12: Salary Predictions from the Addition of "Cisco Certified Network Professional (CCNP)"

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Panel A: Full Sample of Postings								
Original	40631	10.794	0.502	9.632	10.378	10.738	11.207	12.143
CCNP	40631	10.809	0.495	9.654	10.398	10.755	11.218	12.140
Diff	40631	0.015	0.031	-0.292	0.001	0.010	0.025	0.562
Panel B: Only Occupations That Ask for Credential								
Original	22582	10.934	0.503	9.632	10.504	10.946	11.358	12.143
CCNP	22582	10.947	0.496	9.630	10.522	10.962	11.366	12.136
Diff	22582	0.012	0.031	-0.309	-0.001	0.009	0.021	0.549

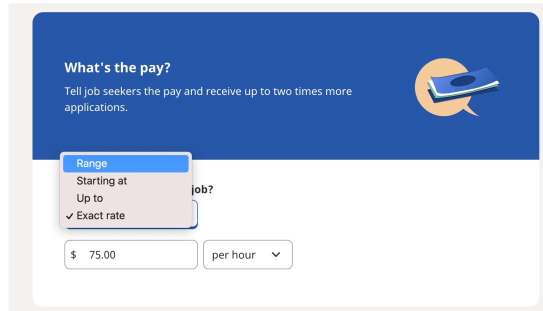
Notes: This table shows the distribution of predicted salaries for a random sample of postings using Greenwich.HR data. Panel A demonstrates the change in the distribution upon adding Cisco Certified Network Professional (CCNP) to the posting. The third row is the difference between the posting with and without the credential. Panel B repeats this exercise with only occupations that request a Cisco Certified Network Professional certification in the first quarter of 2019.

Figure A1: Screenshots of Job Board User Interfaces for Recruiters To Input Salaries

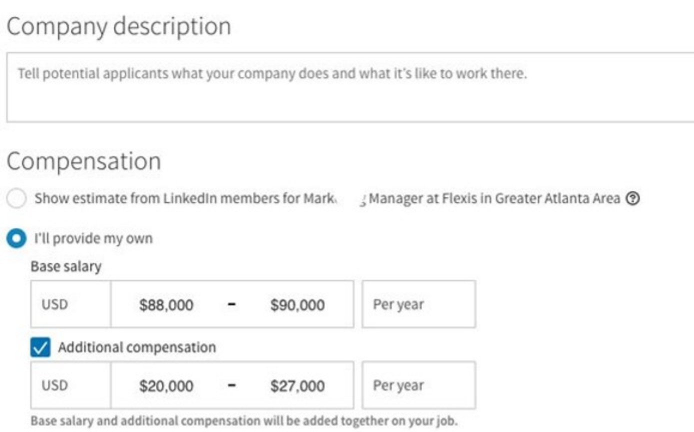
(a) Indeed Posting Screen for Recruiters



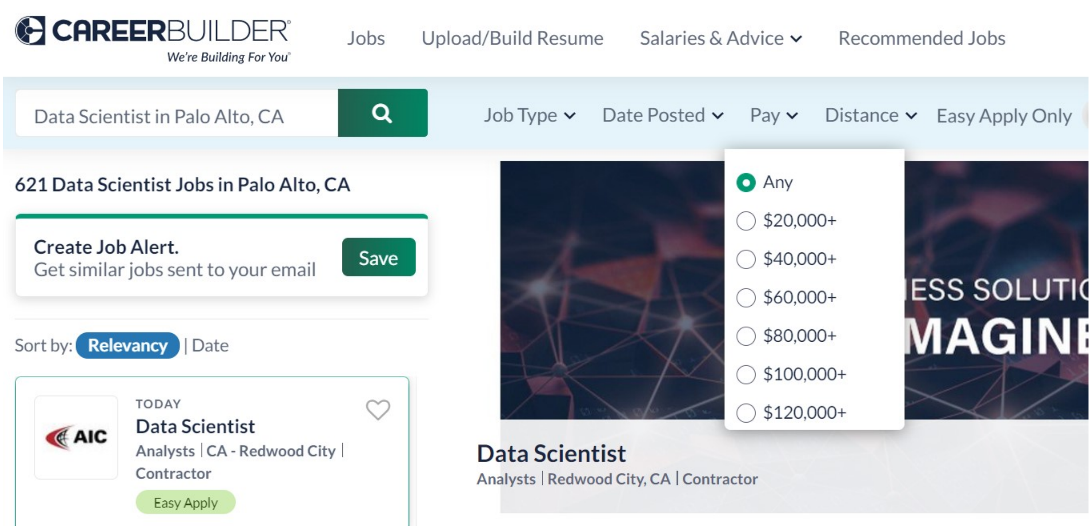
(b) Indeed Options for Recruiters



(c) LinkedIn Compensation Screen for Recruiters



(d) Career Builder Search Portal for Applicants



Notes: This figure demonstrates recruiter side of job posting platforms, which provide the opportunity for recruiters to input salaries. In Panel A, a recruiter is asked the pay for the job. They are incentivized by the statement, "Tell job seekers the pay and receive up to two times more applications." In Panel B, options are displayed. A recruiter can input a range, starting at, up to, or an exact rate. In Panel C, this is the screen on the popular site, LinkedIn. Recruiters are even asked for base salary and additional compensation in separate fields. Finally, in Panel D, you can see the applicant side on another platform, CareerBuilder. The search tool allows applicants to search above a certain pay threshold.

Table A1: Skill Cluster Families

Administration	Human Resources
Agriculture, Horticulture, and the Outdoors	Industry Knowledge
Analysis	Information Technology
Architecture and Construction	Legal
Business	Maintenance, Repair, and Installation
Customer and Client Support	Manufacturing and Production
Design	Marketing and Public Relations
Economics, Policy, and Social Studies	Media and Writing
Education and Training	Personal Care and Services
Energy and Utilities	Public Safety and National Security
Engineering	Religion
Environment	Sales
Finance	Science and Research
Health Care	Supply Chain and Logistics

List of Skill Cluster Families used in Column (3) of Table 2.