The Future is Not Ours to See

Lingfei Wu
Contact: liw105@pitt.edu
The University of Pittsburgh

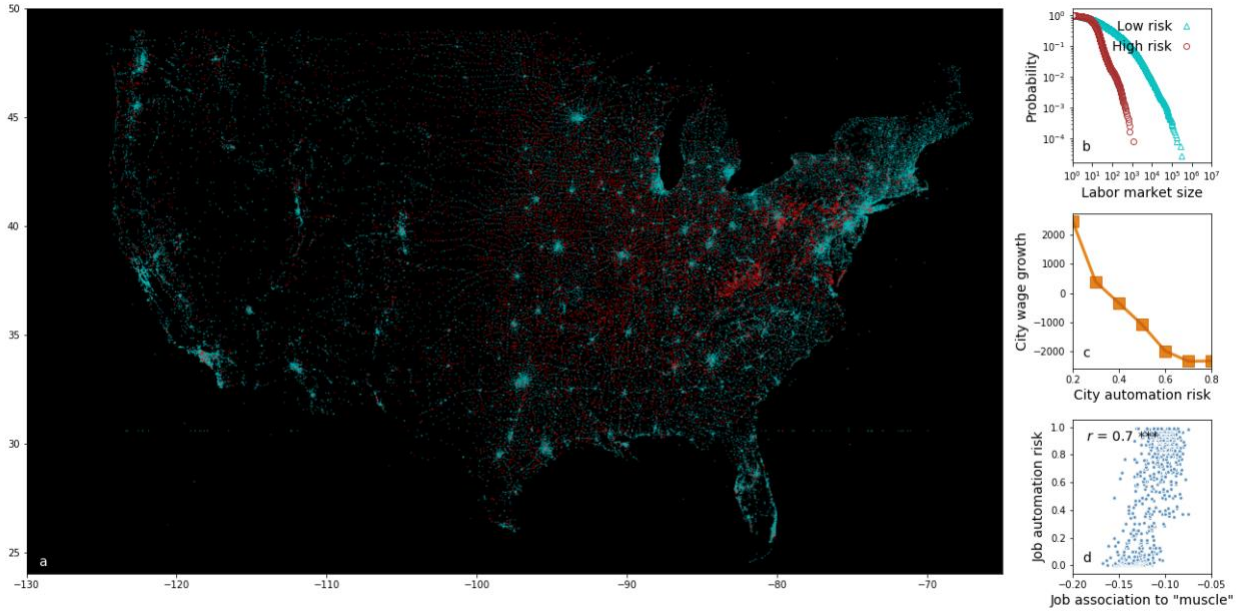

Figure 1. The Future is Not Ours to See. Three datasets are used (see Data for more information). The main source of data is BG, which includes 180,605,633 job advertisements in the U.S. from 2010 to 2019. **a,** Visualizing U.S. locations of high vs. low automation risk. We select 47,990 locations of three or more job openings in BG. For each location, we obtain the local labor market profile by calculating the market share of 1,060 O*NET jobs across ten years from 2010 to 2019. We then obtain the automation-risk scores of locations by selecting the risk scores of jobs from OC and calculating their averaged value, weighted by market shares. Locations are colored by automation risk, brown for high-risk (risk score > 0.65, top 25%) and cyan for low-risk (risk score <= 0.65, bottom 75%). The logarithmic values of labor market size and automation risk score are negatively correlated (Pearson correlation coefficient equals -0.11, P-value < 0.001). Point size is proportional to job market size (see **b** for definition). **b,** Small cities are more vulnerable to automation. We calculate the average number of yearly job openings for 11,998 locations in the high-risk group and 35,992 locations in the low-risk group. We then plot this variable (the x-axis) against its cumulative probability (the y-axis) following the schema of Zipf's law (Zipf, 1935). We plot data points (locations) from high-risk (cyan) group and low-risk group (brown) separately. **c,** Automation risk predicts the decrease in wage. We estimate the median annual wage for the 1,060 jobs using 2018 O*NET data. This information is used to calculate the average yearly wage for each location based on its job market profile. By regressing the average yearly wage against years, we derive the annual growth rate in wage (y-axis) and plot it against the automation-risk scores of locations. These two variables are negatively correlated (Pearson correlation coefficient equals -0.31, P-value < 0.001). We only show binned data in the panel. **d,** Physical jobs are more vulnerable to automation. We estimate the association between jobs and the semantic dimension representing physical efforts

("muscle") using word embeddings (see Method for more information). The association to physical efforts correlates with the automation-risk of jobs (Pearson correlation coefficient equals 0.7, P-value < 0.001).

Significance

This research project responds to the ongoing large-scale replacement of human workers by machines across countries. Motivated by the recent findings that automation has been transformed economic landscape across cities and countries, we explore how small and large cities differ in automation vulnerability. By analyzing more than 180 million job advertisements across 48k locations in the U.S. over the past ten years (2010-2019), we demonstrate that small cities are more vulnerable to automation than large cities. We find that the labor markets in smaller cities are feature by physical occupations, which are easier to automate, and the wage to these jobs was steadily decreasing in the past decade. In contrast, jobs in large cities take more cognitive efforts, harder to automate, and the work pay has been increasing.

Findings

1. Small cities are more vulnerable to automation than large cities. Across the studied 47,990 locations, the Pearson correlation coefficient between the logarithmic values of labor market size and the automation-risk score is -0.11 (P-value < 0.001). The largest labor market in the high-risk group (risk score > 0.65, top 25%) only provides 1000 jobs, whereas its counterpart in the low-risk group (risk score <= 0.65, bottom 75%) provides over 300,000 jobs - three hundred times larger. See Frank et al. (2018) for similar findings.

2. Automation risk predicts a decrease in wage. Across the analyzed locations, the Pearson correlation coefficient between the annual growth rate in wage and the automation-risk score is -0.31 (P-value < 0.001).

3. Physical jobs are more vulnerable to automation than cognitive jobs. Across 1,060 jobs, the Pearson correlation coefficient between job association to physical efforts ("muscle") and the automation-risk score is 0.70 (P-value < 0.001).

Data

*The Occupational Information Network Dataset (O*NET).* O*NET is an online database that contains occupational definitions. It is freely available at https://www.onetcenter.org/database.html. The version used in this study includes the importance scores (ranging from 1 to 5) of 35 skills and 33 knowledge fields that define 966 jobs.

The Burning Glass Dataset (BG). BG is a dataset of job openings. It includes the 180,605,633 advertisements of 1,060 different O*NET jobs across 52,979 locations in the U.S. from 2010 to 2019.

The Occupation Automation Risk Dataset (OC). OC lists 702 O*NET jobs and their risk of automation. These scores are inferred from a training dataset containing 70 O*NET occupations

and their label of "computerizable" (either 0 or 1) assigned by a panel of artificial intelligence experts. This dataset is provided as an appendix in the paper by Frey & Osborne (2013).

Method

Identifying Physical and Cognitive Jobs using Word Embeddings

Mikolov et al. (2013) proposed the *word2vec* model to represent the semantic meanings of words by vectors trained from text data using artificial neural networks. An impressive application of *word2vec* is completing word analogies, such as identifying V(queen) as the vector closest to "V(women) - V(men) + V(king)" in Cosine distance.

After the paper of Mikolov et al., *word2vec* and its variations are widely used to analyze large-scale copra. Several pre-trained word vectors are available to the public, including 300-dimension Google News vectors (https://code.google.com/archive/p/word2vec/), 300-dimension Wikipedia vectors (https://nlp.stanford.edu/projects/glove/), and 200-dimension Twitter vectors (https://nlp.stanford.edu/projects/glove/). Studies on word analogies using these datasets showed that the subtraction between the vectors of antonym pairs gives the universal semantic dimensions. For example, to complete the word analogy mentioned above, the *word2vec* model defines a "feminine" dimension/vector as "V(women) - V(men)" and then searches for the feminine version of "king" along this dimension in the vector space (Mikolov et al., 2013; McGregor et al., 2016).

Pre-trained word vectors (or word embeddings) are providing fruitful social insights outside the filed of machine learning. Caliskan et al. showed that the fraction of female workers within each occupation is strongly correlated with the Cosine distance from the vector representing female to vectors of occupation names (Caliskan et al., 2017). Garg et al. analyzed pre-trained Google News word vectors and found a decrease in gender bias from in the past century (Garg et al., 2018). Kozlowski et al. showed that the dimension of class existed widely in sports, food, music, vehicles, clothes, and names (Kozlowski et al., 2019).

In the current study, we download the 300-dimension Google News vectors (https://code.google.com/archive/p/word2vec/), and construct a "muscle" vector to represent physical efforts by calculating the V(muscle) – V(brain). We then take the 68 job-defining words in O*NET (35 skills and 33 fields of knowledge) and calculate the Cosine similarity from their vectors to the constructed "muscle" vector. We average this value within each job, weighted by their importance scores to the job, to derive the association between 966 O*NET jobs and the "muscle" vector. We find that the studied jobs polarize on this dimension, supporting previous studies on the physical-cognitive polarization of U.S. jobs (Alabdulkareem et al., 2018).

Discussion

Technology tends to create more jobs than it destroys, but this always happened in the long run. Our time is witnessing a massive-scale replacement of human workforce by machine workforce.

This saves human workers from dangerous or tedious jobs on one hand and presents an urgent challenge for our society to adopt automation quickly enough to progress on the other.

Automation is reshaping our economic and social landscapes dramatically. There are fewer jobs and more gigs. Machine substitution is happening fast to deskill human workers, who stuck in the past, holding on to their outdated skills and knowledge. The education needed to work with sophisticated machines is getting more expensive. Worker unions are losing their organization power, as work is decomposed and distributed to tens of millions of workers who do not know each other. They are renamed as contractors to justify a deprivation of welfare under cover of the fancy term sharing economies.

Yes, new job opportunities will be created. However, will these opportunities go back to where they were taken? If the answer is no, and if machines are invented to replace workers who were arranged to work like machines due to a lack of educational resources, where does automation leave them? How to protect individual, low-educational workers, who are more vulnerable than ever, both in the U.S. and elsewhere?

Indeed, we are obligated to think about these questions. The changes brought by automation are penetrating from technical and economics domains into political and cultural realms (Harari, 2015). We have seen the reshape of ideological landscapes as the consequences of the tension between human and machine, or, the tension between people who are taking advantages of automation and people who are suffering from it. An example is the rise of right-wing politics and conservatism in the U.S. and globally. The high-risk and low-risk locations presented in Figure 1 is strikingly similar to the 2016 United States presidential election map (http://nymag.com/intelligencer/2018/03/a-new-2016-election-voting-map-promotes-subtlety.html). And this will only be a signal of a sequence of more disruptive changes to come, if not understood and address timely and effectively.