Workforce Data Initiative Creates Research Hub for Labor Market Data

October 31, 2017
By Monica King

Access to up-to-date workforce data is instrumental in helping government agencies, researchers, and businesses learn about today's rapidly changing labor market conditions to make better decisions. Traditionally, government agencies have produced high-quality labor market data based on nationally representative surveys. But these data are limited in their slower moving nature and aggregation to larger geographic areas.

Workforce Data Initiative (WDI) at the University of Chicago's Center for Data Science & Public Policy is transforming how data on jobs, skills, and training are collaboratively generated and shared with researchers and the public. A recent grant from the Alfred P. Sloan Foundation, along with funding from the JP Morgan Chase Foundation, enabled WDI to create the [Open Skills Research Hub](#), allowing individual researchers to more easily access large private sector datasets of job postings, resumes, and hiring decision, for example. The Research Hub also creates new aggregate public datasets that can provide insight into labor market dynamics, such as changes in occupations and skills, with greater temporal and geographic granularity than what is possible through national statistics.

New Approaches

WDI's innovative approach is to incorporate untapped administrative data from both public sector organizations as well as private companies that provide a constant flow of labor market transactions to capture underlying labor market dynamics.

"Our goal was to figure out the necessary legal, institutional, and technical infrastructure to be able to have a valuable pool of high-frequency and high-granularity public and private data on labor market transactions," Matthew Gee, senior fellow and Co-PI on the project, said. "And to be able to build that infrastructure in a way that a broad set of researchers can access."

Gee and his team of data scientists and software engineers pool and standardize valuable information from job postings, resumes, and applicant profiles. They leveraged data from private sector sources such as CareerBuilder in addition to public sector sources like the National Labor Exchange and have worked out collaborative legal agreements that allow for approved researchers from outside the University of Chicago to gain controlled access to these normalized private data for their own research.

In addition to the access controlled private data in the Research Hub, the publicly available Research Hub datasets, which are divided into quarters at the metro area level, open up new opportunities for discovery by researchers and policymakers. "There are a lot of unanswered yet essential questions about how people signal their

competencies and abilities, how firms organize their labor, and how individuals invest in their own human capital," said Gee. "These are all important fundamental questions that require an understanding of the labor market not just at a high level in broad strokes but at an increased granular level of skill and competency."

Data Challenges

The team faced a number of challenges in working with the private sector data sources. First, the company data come in largely unstructured form. In order to correctly assign occupation codes, for example, the team's data scientists use natural language processing and machine learning techniques to make sense of the raw data.

"Because we're getting data from different sources, and each of them is structured in different ways," Tristan Crockett, a software engineer on the team, said, "coming up with analysis that works well for all of the data we have is a challenge."

Another challenge is related to the representativeness of the private sector data. Eddie Lin, a data scientist on the team, explained, "You probably won't see a truck driver job posting on LinkedIn data." To assess just how representative (or not) the Research Hub are, [Lin compares the WDI data](#) to representative labor market data from the Bureau of Labor Statistics. The team also plans to leverage data from many additional sources in order to reduce the potential selection bias of their data product.

Next Steps

Gee and his team are continuing to pool more data into the Research Hub, a process that first begins with building relationships with new companies and data providers. WDI is also continuing to host monthly working group calls with users of the data, both as an opportunity for the team to give updates about the data and also for researchers to share the types of questions they are hoping to answer with the data.

"The community of practice around the use of the data is just as valuable as the data itself," Gee said, "because it ensures that we're going out to negotiate the right types of data sharing agreements and that the core team is setting up the researchers for success."

Eddie Lin will present on this project at the upcoming [ADRF Conference](#) on November 14, 2017.