Education Data Science Fellowship
Strategic Data Project at Harvard University

Overview

The Strategic Data Project (SDP), an initiative of the Center for Education Policy Research at Harvard University, is offering a one-year Education Data Science Fellowship starting in the fall of 2021. The fellow will work at a US Federal Agency in Washington, DC and engage in transformative data science and research projects using federal education data.

Education research leaders are working to realize the potential of previously untapped data collected across multiple federal agencies. To carry out the work of bringing together new data sets and tools and publishing novel insights, SDP will host the inaugural cohort of data science fellows to work with federally collected education data and to connect those data across departments to generate new insights about the health, well-being, and achievement of students. The projects fellows may execute will have tremendous impacts on the sector. Fellows may be involved in reinventing and testing indicators of family poverty or helping to accelerate the timeline to identify research insights using natural language processing of formal and informal research sources. Please see the attached project list for examples of the work fellows may be involved in.

The Education Data Science Fellow will have the opportunity to work with Harvard faculty advisors focused on education policy research, as well as research and data leaders in the department. Additionally, they will engage in ongoing research projects and will participate in the design of new studies utilizing federal data sets at a summit of faculty, fellows, and researchers that kicks off the fellowship program. Finally, fellows will have the opportunity to access Strategic Data Project training opportunities as desired.

Qualifications

Applicants must have completed their PhD in a relevant discipline and have at least five to ten years of relevant work experience, including a successful work history of designing and implementing large research projects in US public education. Applicants must have experience with large administrative data sets in education, sampling methods and primary data collection processes; in depth knowledge of education policy issues; experience with data processing, archiving, analysis and report writing, including statistical analysis; proficiency in major social science databases and statistical software packages; evidence of leadership and management skills. *Applicants must be citizens or permanent residents of the United States and have the ability to obtain a federal security clearance.*

The skills needed for each project will vary and the Data Science Fellow will match to projects based on the skills they bring. Some of the high priority skills include:

- Ability to extract, summarize and transform data using a relational database management system (e.g. SQL, Oracle, etc.)

- Proficiency in statistical analysis using R or Python. Proficiency using scientific computing packages (e.g. Numpy, Pandas, SciPy)
- Experience in parametric and non-parametric analysis (e.g. decision tree, multivariate analyses (e.g. principal component analysis, clustering algorithms, neural networks)
- Experience in Natural Language Processing feature engineering and modeling including processing of large text collections with standard NLP tools for parsing, entity extraction, part of speech tagging, topic discovery and classification (such as sentiment analysis), and natural language understanding.
- Domain expertise with educational theory (general and special education) and algorithmic approaches to common learning challenges for students with and/or without disabilities (e.g. Learning Analytics, Educational Data mining, Natural Language Processing)

Application Process and Deadline
Applicants should submit the following materials as attachments in a single email to sdprecruitment@gse.harvard.edu. Completed application packages will be reviewed on a rolling basis and candidates will be contacted for interviews.

Applications must include a:
- Curriculum Vitae
- Cover letter detailing your qualifications as they pertain to the job description (2-3 pages)
- Please submit a code repository that demonstrates as many of the following items:
    - Your ability to transform unstructured and complex data
    - Your ability to apply robust statistical techniques (especially non-parametric and unsupervised methods)
    - Your ability to present results in a user-friendly format
    - Your ability to create clean and reproducible code
- A one-page statement of interest discussing which project you would pursue from the attached Federal Project List (*see next page).*
- 3 confidential letters of recommendation sent directly from the references to sdprecruitment@gse.harvard.edu

Salary and Benefits
The stipend for the Data Science Fellow is between $90,000-$150,000 and commensurate with experience. It provides modest health insurance coverage.

Harvard University is an equal opportunity/affirmative action employer.

Education Data Science Fellowship: Federal Project List

These are examples of the types of projects that fellows may be involved in and the skillsets that are required; final projects will be assigned in consultation with selected fellows and their administrative sponsor.

Submit a one-page statement of interest discussing which project would you pursue, and why. What questions do you have about this project? How might you approach the work?

Using NLP to Identify Leading Indicators of Education Research

Conventional research meta-analysis is time-intensive and the practical utility of this analysis is limited by delays that result from the time required to do high-quality meta-analysis; by the time these analyses are published the practices being examined are often well-established – whether or not there is evidence to support them. This project will accelerate the timeline to identify research insights using natural language processing of formal and informal research sources. The project will identify themes, methods, and promising approaches for practitioners and researchers that could be used in public-facing web applications and to inform published studies and further in-depth analyses.

Modeling Process Data for Insights into Assessment Design

Clickstream data collected by computer-based assessment platforms can be used to model the cognitive and behavioral processes of respondents. While these techniques are commonly used in education technologies, there has been a limited amount of work in using these methods to inform assessment design. This project would use this "process data" to model student interaction with different question types in concert with response data to provide deeper insights into item design efficacy. Extension to broader constructs within psychometric theory is also a possible outcome.

Reinventing & Testing Indicators of Family Poverty

The number of students receiving free and reduced-price lunch has long been used as an indicator of poverty, but this measure is becoming increasingly invalid as policy around free lunch distribution and reporting changes. New geo-spatial and other measures that combine multiple data sets are being developed to provide a more accurate understanding of family economic status. This project would support these efforts to conduct simulations of these new approaches and compare them to the current approach to identify promising approaches and limitations.

Streamlining IPEDS Campus Data Collection & Inferences

The Integrated Postsecondary Education Data System (IPEDS) provides student enrollment, completion, financial aid, and other information for every college and university in the United States that receives federal financial aid. The data elements included in the IPEDS collection were identified by stakeholders interested in educational processes and outcomes. Collecting this data is a significant effort for each university, and it is not clear if all data elements provide useful information. This project will analyze IPEDS data and identify data elements that have low variability, have high collinearity with other variables, or other characteristics that make them unnecessary. An interesting implication of this project is the implications of these inferences, especially as they contrast to common beliefs about educational institutions. The project may also uncover areas where additional data could be recommended.

Skills required

- data engineering, data simulation
- statistical analysis
- visualization and reporting
- higher education administrative data knowledge required

Modernizing IES Administrative Data Collection Formats

IES maintains the two largest sources of information about K12 and higher education administrative data with the Integrated Postsecondary Education Data System (IPEDS) and Common Core of Data (CCD) collections. These data in their current structures, information about school and university characteristics, enrollment counts, and aggregated statistics on resources, have been collected for more than 25 years. Data are stored in formats that were determined based upon data processing needs and have not been evaluated to determine limits on the types of analysis that can be performed. Specifically, data are stored in SQL tables that can require significant computing power and time to execute queries. This project will evaluate the structure of these databases and prototype enhancements to data storage that might enable more contemporary approaches to data analytics and also may result in cost reductions through the use of more advanced infrastructure.

Skills required

- data engineering, data simulation
- statistical analysis
- visualization and reporting
- higher education administrative data knowledge required

Improving Quality Control Rules & Approaches for Administrative Data

IES maintains the two largest sources of information about K12 and higher education administrative data with the Integrated Postsecondary Education Data System (IPEDS) and Common Core of Data (CCD) collections These data in their current structures, information about school and university characteristics, enrollment counts, and aggregated statistics on resources, have been collected for more than 25 years. To ensure that data are accurate and no obvious errors have been made, IES runs a series of business rules to identify anomalous and possibly erroneous data. These rules have been developed independently over time for each collection. This project will analyze these rules and synthesize across the collections for optimization and consistency. The potential for school-level QA analytics will also be explored. The project will involve the simulation of results and report recommendations.

Skills required

- data engineering, data simulation
- statistical analysis
- visualization and reporting
- higher education or K12 administrative data knowledge required