This posting is for a summer position working with the Automated History Archive, an initiative to automate the conversion of image scans of historical quantitative publications into classified, machine-readable data models. Many of the biggest challenges that our society faces have their roots in the past, and history can provide fundamental insights into their causes and potential solutions. While there have been major advances in cataloging past knowledge, vast amounts of historical quantitative data that could shed light on important issues, such as inequality, social upheaval, and economic growth, remain locked in hard copy due to prohibitive curation costs. Automation and open collaboration can unlock vast disaggregated data, spurring research on a diverse array of important social science questions.

A successful automation pipeline for converting raw image scans into classified data models requires integrating computer vision tools that can recognize highly irregular data structures in the raw images with machine learning techniques for classifying digitized table fragments. Building on our initial experimentation with automating the digitization and classification of complex historical publications, we will be using tools from computer vision and machine learning to automate the conversion of historical quantitative documents into classified, machine-readable datasets on a large scale. The output will be deposited in a collaborative data platform.

The student will work on algorithms for assembling digitized table fragments into classified data models. Excellent programming and problem-solving skills, as well as some background in machine learning, are required. Both undergraduate and graduate students are welcome to apply. The full-time position is for eight to ten weeks during the summer (flexible dates). The student will sit with our other team members at the Institute for Quantitative Social Science, which offers a rich community of researchers working on data science problems with applications to the social sciences. The position provides an excellent opportunity to hone data science skills by applying them to important social questions.

Interested candidates should send a CV, unofficial transcript, and cover letter describing their experience with data science to the project lead, Professor Melissa Dell (melissadell@fas.harvard.edu).