Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings

José Bayoán Santiago Calderón
Bureau of Economic Analysis
Jose.Santiago-Calderon@bea.gov

Dylan Grassier
Bureau of Economic Analysis

The current global efforts for revising the System of National Accounts have identified valuing and recording data as a high-priority area. The emergence and growth of data-enhanced and data-enabled businesses such as online platforms have contributed to the status of data as a vital asset in modern economies. These ongoing efforts include developing a taxonomy to identify the scope of data assets, measure them, and incorporate them into the national accounts (OECD 2021). Current guidance encourages national statistical offices to explore and share potential methods, estimates, and conceptual bases used in these efforts.

The U.S. Bureau of Economic Analysis has developed a machine learning application to estimate time-use factors for data-relevant activities per occupation using the text of online job advertisements. The occupations' time-use factors can be decomposed into two components: (1) the average share of time allocated to data-relevant tasks and (2) the share of employees engaging in those job activities. Using Burning Glass Technologies (BGT) job advertisement data, we identify which skills in the BGT taxonomy are "data-related" as relevant to data entry, storage, analysis, or management. We identify occupations with the highest rate of job openings containing "data-related" skills. The top occupations are denoted as "known" data-intensive occupations and serve as "landmark" occupations (e.g., statisticians). A doc2vec model is trained on the job advertisement text for each occupation to obtain a high-dimensional representation of what the occupation-level job postings convey. Using this numerical representation, we can obtain occupation-level pair-wise distances to measure how "close" or similar an occupation is to the "known" data-intensive occupations. The product of the similarity to a landmark measure and the ratio of job openings with identified "data-relevant" activities serves as the proxy of the occupation-level time-use factor. The BGT data also enables additional paths to adjust for overlap with other intellectual property products currently captured in the U.S. national accounts as capital formation, including own-account software and own-account R&D.

Using this method, we obtain cost-based estimates of annual investment in data-relevant activities in current dollars for 2010 – 2019, during which it grew from $88.3 billion to $130.8 billion, which yields an average annual growth of 4.4 percent. In addition to the estimates for capital formation, we present net stock estimates for data as an asset. We explore the impact of accounting for capital formation in data at industry levels and on other national aggregate accounts, such as software, aggregate IPPs, and value-added of the business sector.

If incorporated into the U.S. national accounts, data as an asset would amount to expanding the production boundary as part of the own-account software and databases category of IPPs, which currently excludes the value of the embedded information content (i.e., data) for own-account.