IARIW-ESCoE Conference
“Measuring Intangible Assets and Their Contribution to Growth”

# Speaking the Same Language: a Machine Learning Approach to Classify Burning Glass Skills

*Luca Marcolin, OECD, luca.marcolin@oecd.org*

*Julie Lassébie, OECD, julie.lassebie@oecd.org*

*Marieke Vandeweyer, OECD, marieke.vandeweyer@oecd.org*

*Benjamin Vignal, ENSAE, benjamin.vignal@ensae.fr*

The measurement of changes in skill requirements in jobs and occupations has been a longstanding challenge, due to the limited availability of granular data across time and space. Recent large databases of online job postings may offer a solution, as they assemble information on a vast set of job characteristics, including skills and education requirements. Yet the number of distinct skills listed in such datasets can be very large and pose several analytical challenges, from the treatment of synonyms to the need to sensibly group skills for inference.

The present study proposes an original approach to reduce the dimensionality of the skill information contained in one such dataset of online job requirements, the Burning Glass Technologies dataset. It does so by classifying the different skills appearing in Burning Glass data into a pre-existing skill taxonomy based on the skill’s meaning or definition. Because of the sheer number of unique skill keywords in Burning Glass (approximatley 17 000 in our sample), a manual classification was ruled out. We propose instead a semi-supervised machine learning approach that produces an automatic classification of skills into the taxonmy’s broader categories. The approach builds on BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art algorithm published recently by researchers at Google AI Language.

The final taxonomy largely builds on pre-existing ones, which were developed and validated by labour market and education experts, and have been widely used by policy-makers and statistical agencies. In particular, we exploit the U.S. Occupational Information Network (O*NET) database, with its detailed information on skills requirements in occupations and its organisation of skills in a clear hierarchical structure. We complement these data with information from the European Skills, Competences, Qualifications and Occupations (ESCO) database in those areas where O*NET is not sufficiently detailed. The resulting taxonomy contains 61 categories. Its stability of structure over time and mutually exclusive categories facilitate the measurement and analysis of human capital in the labour market.

The approach we propose classifies the 17 331 skills keywords contained in the original Burning Glass data into the 61 skill categories of our O*NET-based taxonomy with a satisfactory accuracy that does as well as a manual allocation of a sub-sample of skills. We conclude that our approach is a suitable tool to classify a large amount of data on skill requirements while avoiding an extremely time-consuming and onerous manual task.

As a result of the classification exercise, among the most populated skills categories are “Medicine and Dentistry” and “Biology” , which partially reflects the high frequency of health-related job postings in the Burning Glass database, and their propensity to include very specific skill requirements. Conversely, skills relating to human values or personality traits are very rare, possibly because they are implicit and are not clearly stated in the job advertisement in the first place.

Lastly, we validate the results of the classification exercise in two ways. First, we compare the skill requirements of occupations as listed in the O*NET database and those based on the re-classified Burning Glass skill information. We find that the two data sources provide a coherent picture, at least for the occupations were the representativeness of Burning Glass data has been corroborated. Second, we use the classification to reproduce patterns of skill demand and skill-wage elasticities derived in previous studies. For instance, we can reproduce selected evidence on skill requirements and wage returns to skills in Deming and Kahn (2018) and in the spirit of Autor and Handel (2013), and find intuitive patterns of skill requirements by digital sectors, as in Grundke et al. (2018). Overall our work contributes to the measurement of skill requirements and changes thereof using microdata, as recently summarised by Handel (2020). The data as we structure them can then be used to assess the contribution of changes in skill requirements to a number of relevant economic phenomena, such as the relative importance of cognitive and non-cognitive skills, the changing nature of jobs, or the contribution of skills to productivity growth.