Leveraging natural language processing to infer the skill relevance in occupations

MEMBER STATES WORKING GROUP ON ESCO

10th November 2022

Fabio Manca - Big Data Coordinator

OECD – Directorate for Employment, Labour and Social Affairs

Context

Online Job Postings (OJPs) contain detailed information about the skills, knowledge areas, tasks and technologies required for a job

But

1. They do not rank ‘skills’ by importance or relevance
2. Some of the skills are not mentioned (implicit) even if they are potentially very important (cooking skills for a cook)
3. Simple ‘frequency based’ analyses result in biases when assessing (quantitatively) shortages in labour markets

Leveraging semantic analysis applied to OJPs

Word embedding algorithms (such as Word2vec) are a well-known set of Natural Language Processing (NLP) algorithms.

These algorithms function by **creating a mapping between words and their meaning** (semantics) by turning them into a mathematical representation (**word vectors**) that can be understood/analysed by computers in a **'graph'**

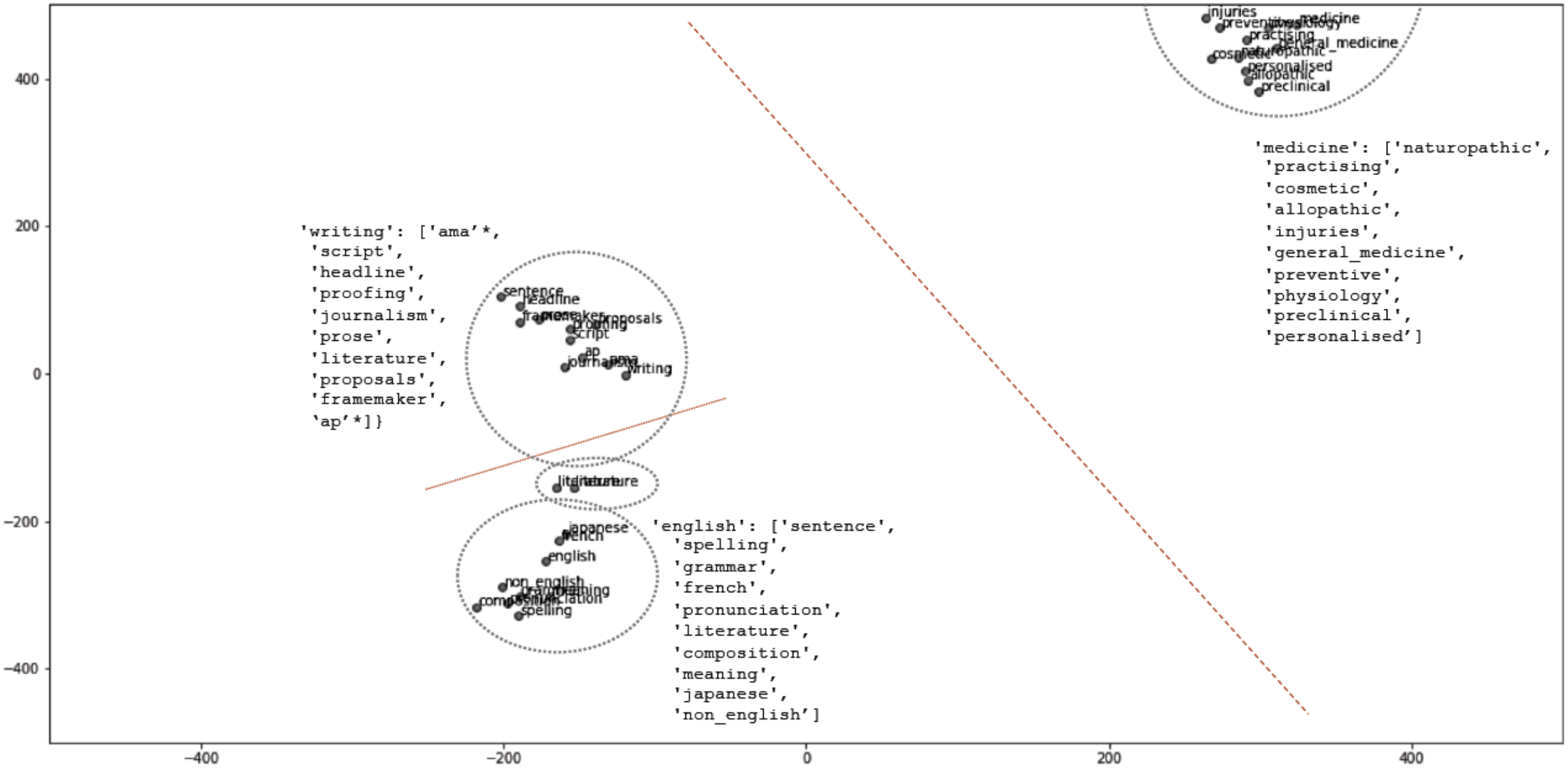
This allow to do vector calculations that retain "semantic meaning" such as:

vec("writing") + vec("python")= vec("programming")

The analysis allows to:

1. **Infer the 'relevance' of skills for any given occupation (through semantic relationships)**
2. **Create matrix of 'occupation similarities'** (based on skill requirements and relevance) that are functional to assess potential career moves based on skill requirements

‘Graph’: the mathematical representation of semantic linkages between words



* ama = AMA Style Guide Writing
ap = AP Style Journalism

Semantic Skill Bundle Matrix (SSBM)

We calculate the (cosine) similarity between each “word vector” (Word2VEC) and any given “occupation vector” (Doc2VEC)

Web Designer		Marketing Manager		Etc…
Web Design	0.73	Online Marketing	0.57	…
Graphic And Visual Design	0.55	General Marketing	0.52	
User Interface And User Experience	0.55	Marketing Strategy	0.50	
Digital Design	0.55	Web Analytics	0.49	
Javascript And Jquery	0.55	Media Strategy And Planning	0.47	
Animation And Game Design	0.53	Content Development And Manageme	0.45	
… continues…		…continues...		

This is a powerful tool, as it replicates the structure (occupations vs skills) of ONET.

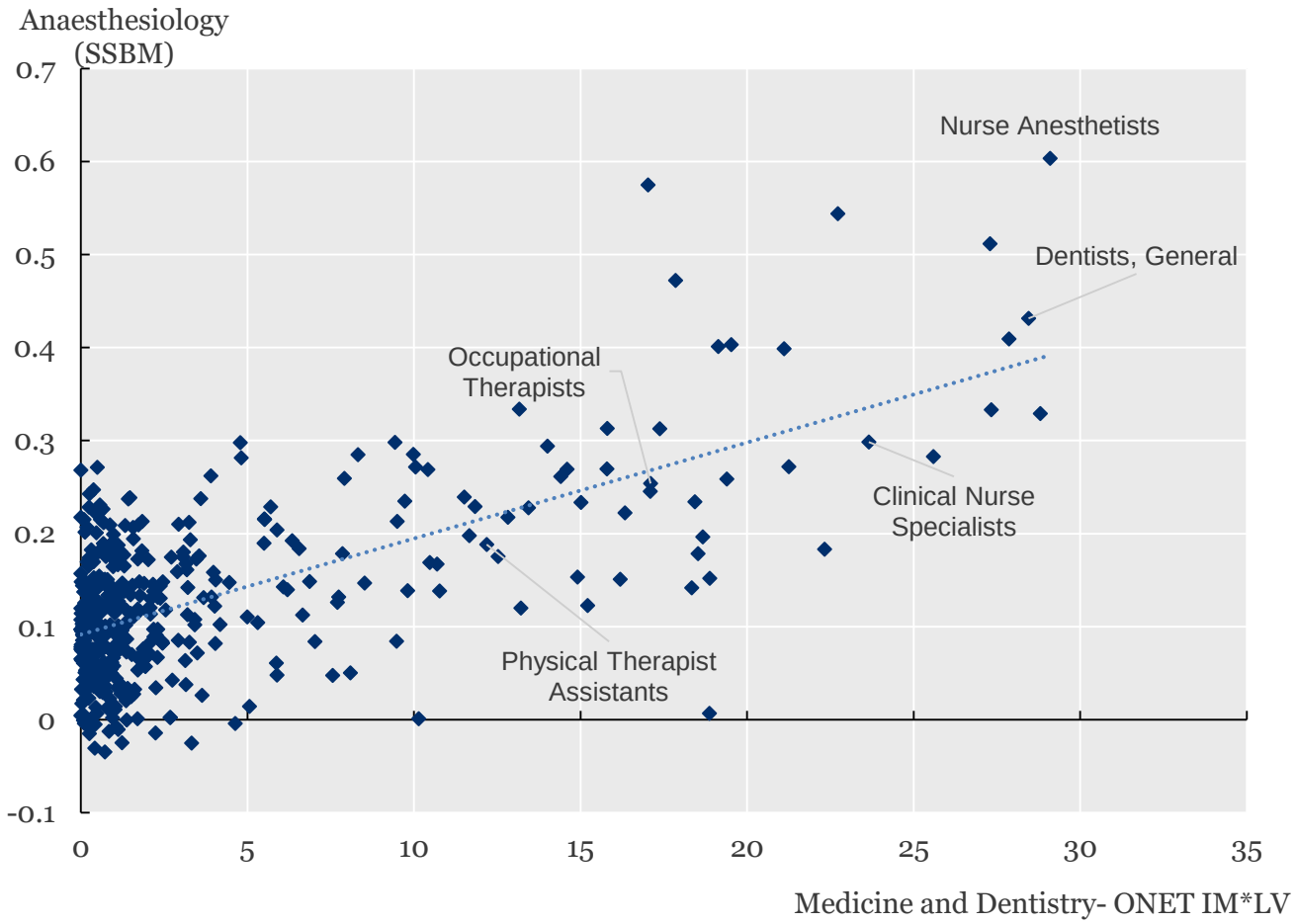
Values in the SSBM represent the ‘relevance’ of each skill for any given occupation

600+ occupations x 1400+ skills

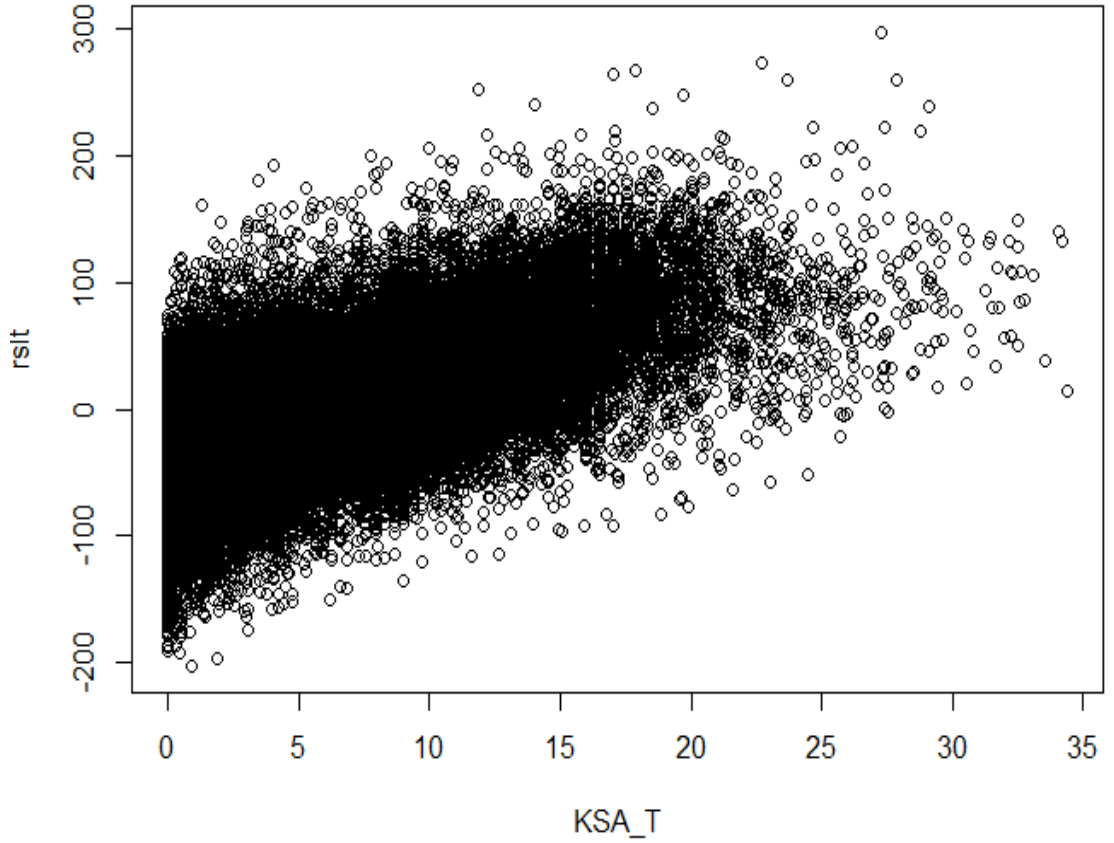
(10x larger than ONET)

Validation of the SSBM: testing against O*NET

Occupation-based correlation SSBM-ONET



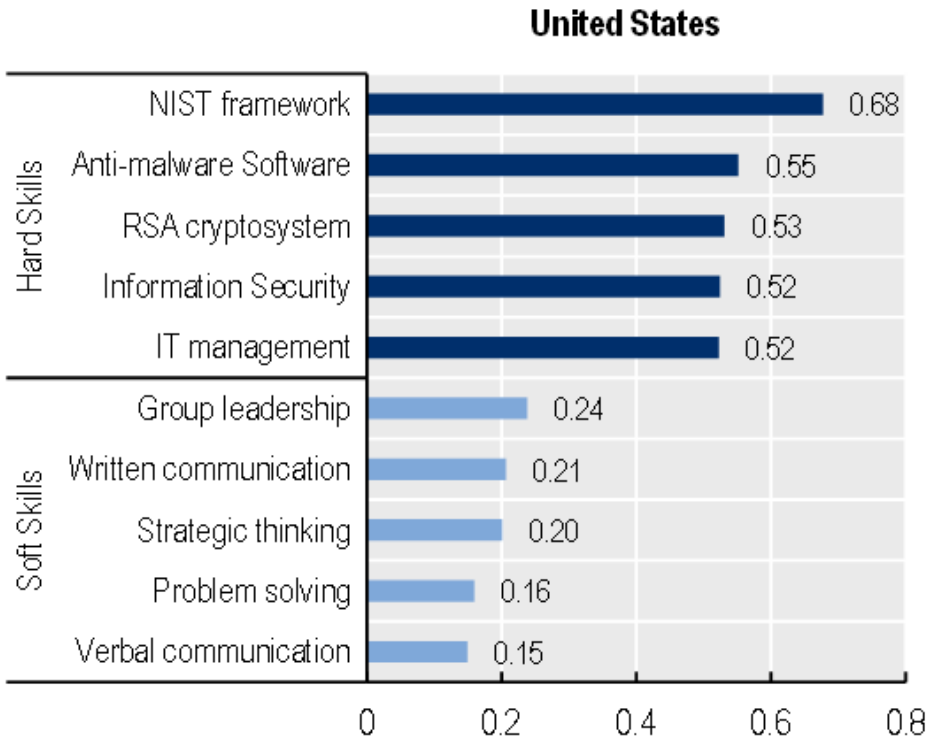
Global correlation SSBM-ONET



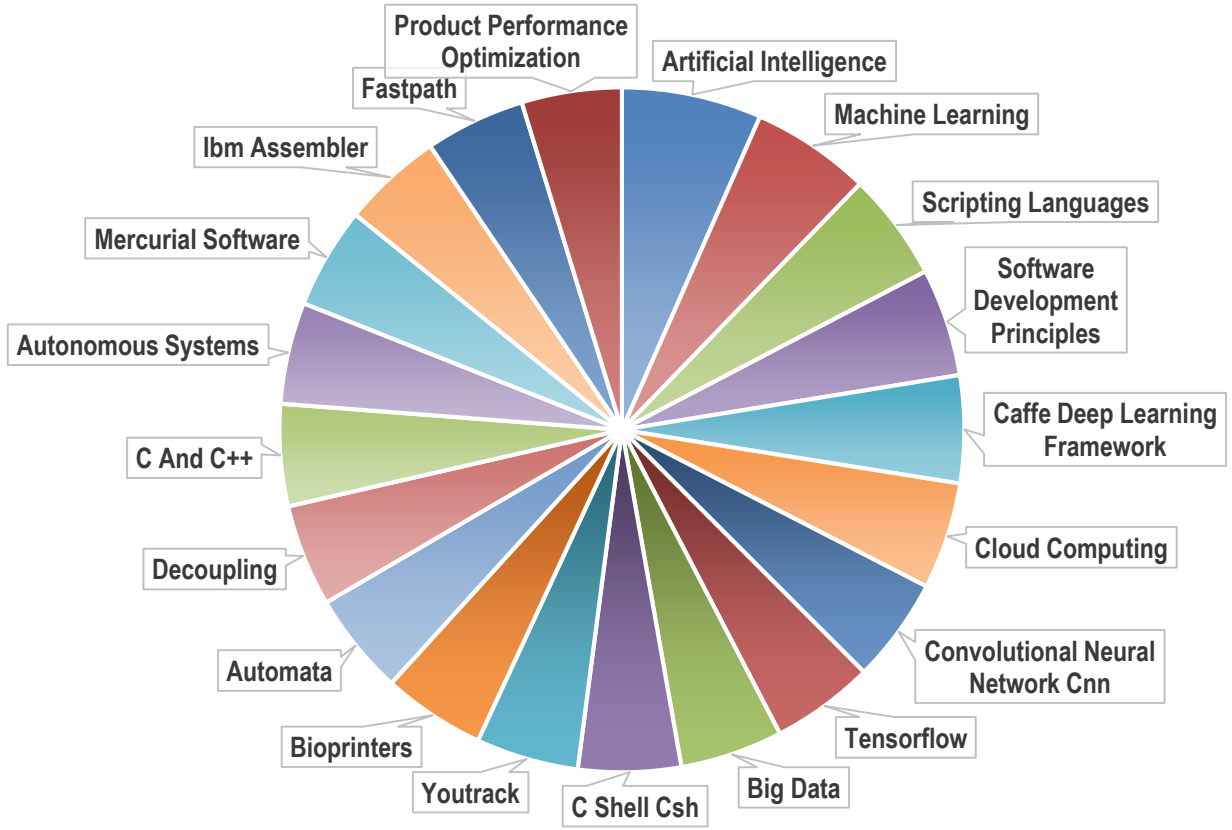
Applications of the SSBM to skill analysis

Granular skill profiles at the occupation level (by skill relevance)

Cyber-security officers

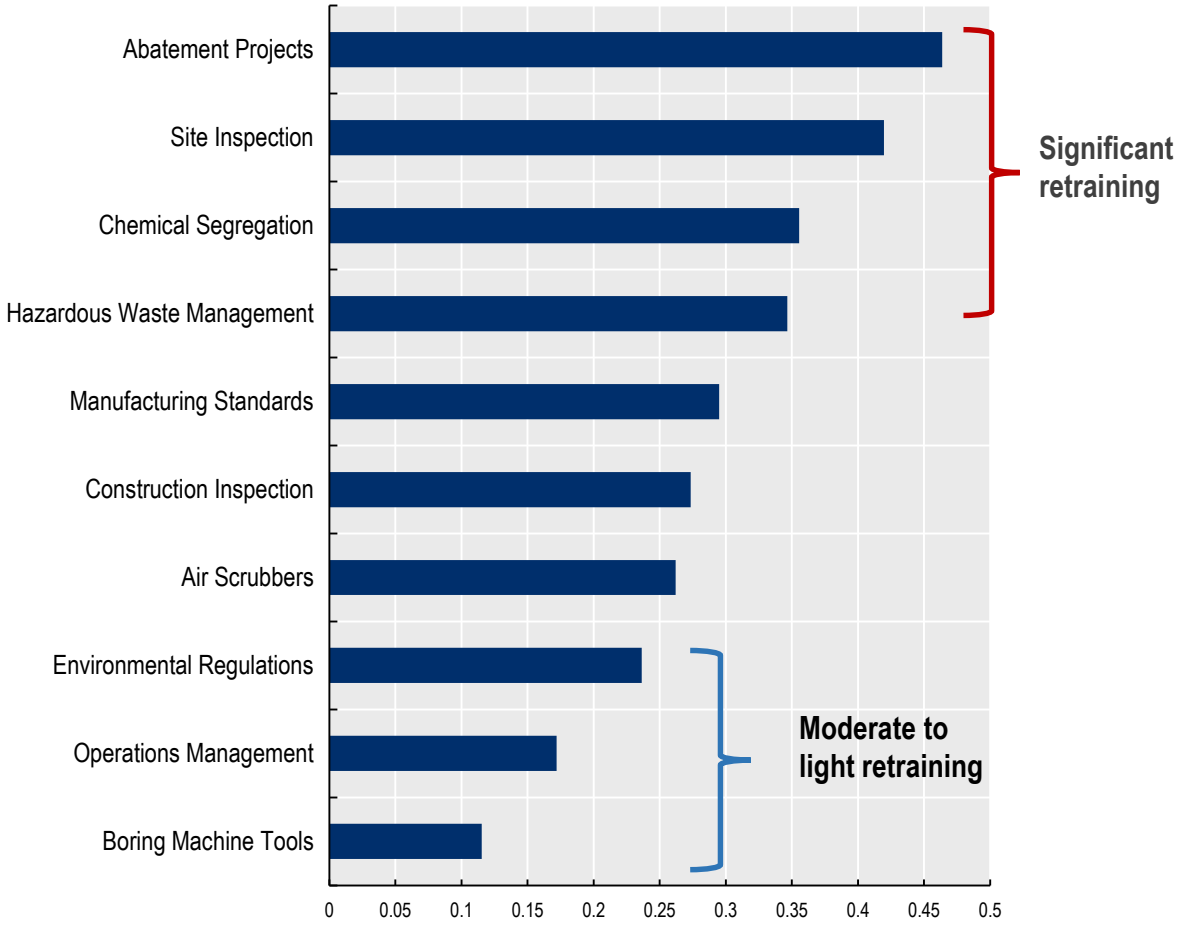


Correlation between skill demands (across occupations)



Retraining pathways from brown to clean/green occupations

Retraining effort from Miners to Hazardous Material Workers



Conclusions

- OJPs have several advantages over more traditional LM statistics, but lack information on the relevance of the skills for the occupation
- Word embedding algorithms can be used to infer the ‘relevance’ of skills in occupations by maintaining a great level of granularity
- This approach aligns with other (much more costly and more ‘static’) data sets like O*NET
- The data on skill relevance can be used to analyse overall shortages in labour markets, skill relationships across occupations and to build retraining pathways from occupations in decline to jobs in high demand

Fabio.manca@oecd.org