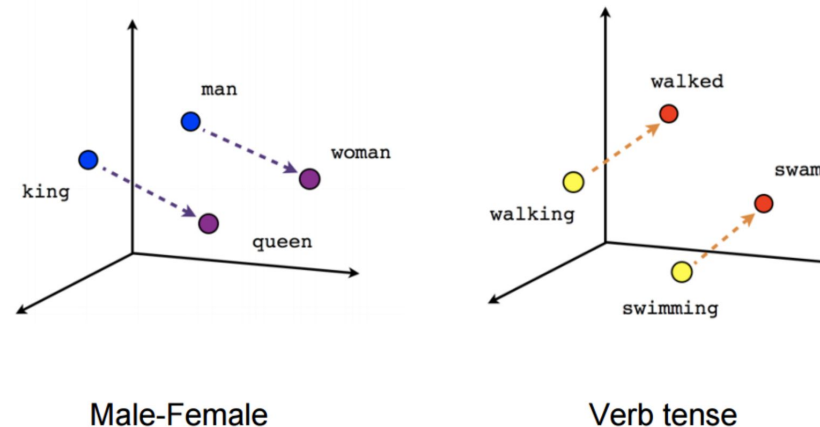




## **Creation of granular skills profiles and skill relevance scores**

# Creation of granular skills profiles and skill relevance scores

- > **Skill keywords in OJPs are useful to make inference about skill demand in labour markets and skill relevance.**
  - > Skill frequencies do not offer a reliable relevance measure since widely common skills (e.g. communication) will drive the results.
  - > We need a robust approach to signal key skills for an occupation from the demand perspective, without relying on frequencies.
- > **Tools from Natural Language Processing (NLP), such as vector space (or embedding) models help.**
  - > Use the word's context to derive its meaning and create n-dimensional vectors to represent that meaning.
  - > Vectors may be plotted as coordinates in a high-dimensional space □ Each dimension stands for a certain context and vector coordinates refer to the count of this context.
  - > We can use this method to analyse interaction among skills and occupations.



- > **Vector representations are suitable for arithmetic operations that assess meaningful connections between vectors.**
  - > Similarity measures for comparing pairs of occupations or skills, as well as any occupation-skill pair.

# Phase I, Output II - Creation of granular skills profiles and skill relevance scores

Steps for creating the vector representations and derive occupation skill profiles

## 1. Preparation of raw data (based on LC skills dataset)

1. Defining the period, geographic and occupation level for the analysis
2. Create a list for each occupation containing the skills required in job postings linked to that occupation.

## 2. Creation of a vector-based text representation of occupations and skills

1. A model for creating vector-based text representations of both skills and occupations: Doc2Vec
2. Creating a customised routine for selecting the best model hyperparameters
  1. Cluster occupation-level vectors to replicate broader occupational categories
  2. Define a metric to compare actual classification with clustering results.
  3. Use the best model to generate occupation and skill vector dataset

## 3. Producing granular skills profiles based on skill relevance scores

1. Skill similarity as a proxy of skill relevance for occupations
2. Skill profiles for selected occupations

# 1. Preparation of raw data (based on LC skills dataset)

## Data sources:

- > **LC job postings database:** Job posting ID and NOC classification.
- > **LC skills dataset:** Skills extracted from job posting descriptions. Linked through a unique posting ID. More than 30k skills in the taxonomy

**Objective:** Create a file that describes NOC occupations based on skill demand.

## Key steps:

2. Defining the period, geography and occupation levels of analysis:
  - > Granular occupations □ A longer list of occupations benefits the model with more examples for training.
  - > Indicators are more reliable when calculated at a national level (more data). Usually a one-year analysis.
3. Cleaning procedures:
  - > Discard less frequent skills (10<sup>th</sup> percentile) to reduce the effect of uncommon skills.
  - > Job postings with a long skill list (95th percentile – 30+ skills approx.). It is unlikely that one job posting could contain that many skills (likely related to scraping problems).

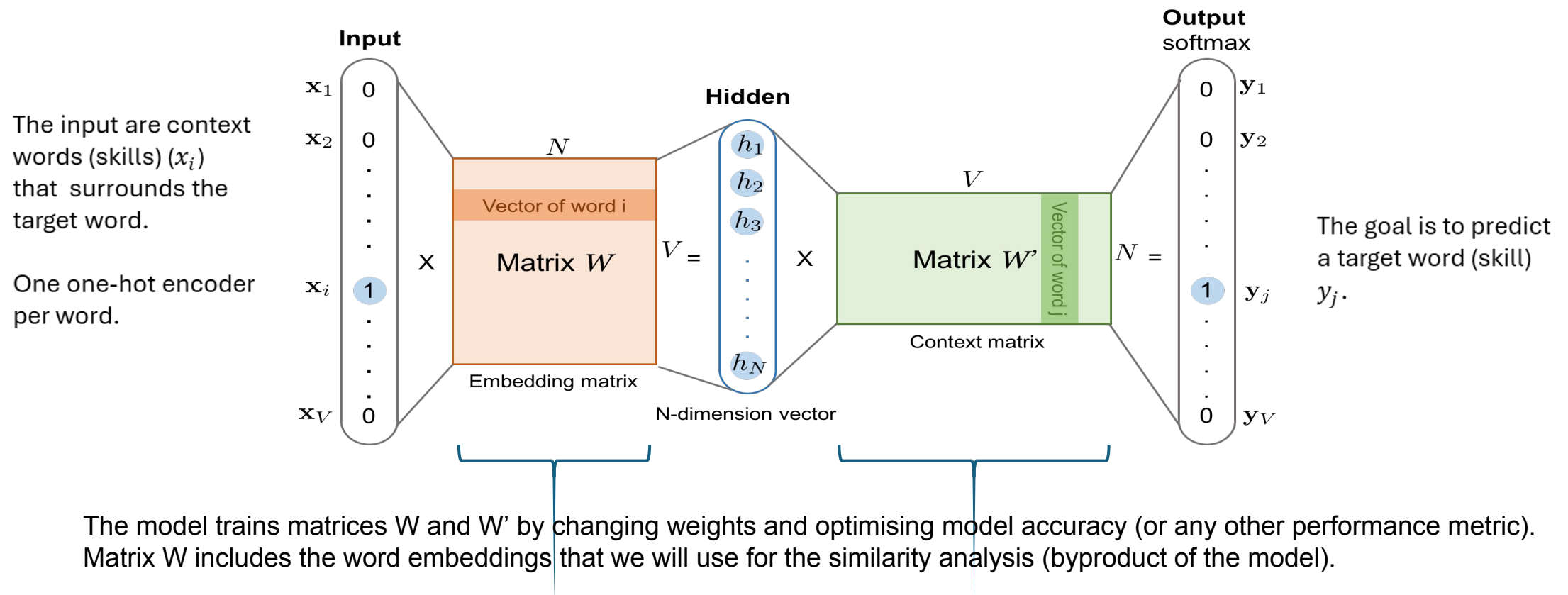
# 1. Preparation of raw data (based on LC skills dataset)

3. Compiling a document for each occupation containing the complete list of required skills:
- > Each skill is pre-processed to remove punctuation, stop words, special characters, etc.
  - > Unique tag for each occupation to enable the model to generate occupation-specific vectors (document vectors)
  - > Skills may appear multiple times within a single occupation.

ID	NOC 5D NAME	SKILLS
0	Financial managers	['Financial Controls', 'Internal Controls', 'Management', 'Interpersonal Communications', 'Verbal <b>Communication</b> ', 'English Language', 'Report Writing', 'Financial Data', 'Operations', 'Detail Oriented']
1	Sales and account representatives - wholesale trade (non-technical)	['Management']
2	Retail salespersons and visual merchandisers	['Customer Service', ' <b>Communication</b> ', 'Merchandising', 'Sales', 'Problem Solving']
3	Transport truck drivers	['Safety Standards', 'Management', 'Waste Collection', 'Hydraulics', 'Waste Removal', 'Inventory Management']
4	Specialized livestock workers and farm machinery operators	['Microsoft Word', 'Self-Motivation', 'Data Collection', 'Embryology', ' <b>Communication</b> ', 'Forestry', 'Time Management', 'Operations Management', 'Machine Operation', 'Valid Driver's License', 'Record Keeping', 'Microsoft Excel', 'Quality Assurance', 'Pesticide Applicator License', 'Irrigation (Landscaping And Agriculture)', 'Problem Solving', 'Pesticides', 'Maintenance Scheduling']
5	Elementary and secondary school teacher assistants	['Individualized Education Programs (IEP)', 'Management', 'Cooperation', ' <b>Communication</b> ', 'Life Skills Development', 'Hygiene', 'Interpersonal Communications', 'Medication Administration', 'Lifting Ability', 'Computer Networks', 'Sales', 'Sports Equipment', 'Toileting', 'Writing', 'Fine Motor Skills']

## 2. Creation of a vector-based text representation of occupations and skills

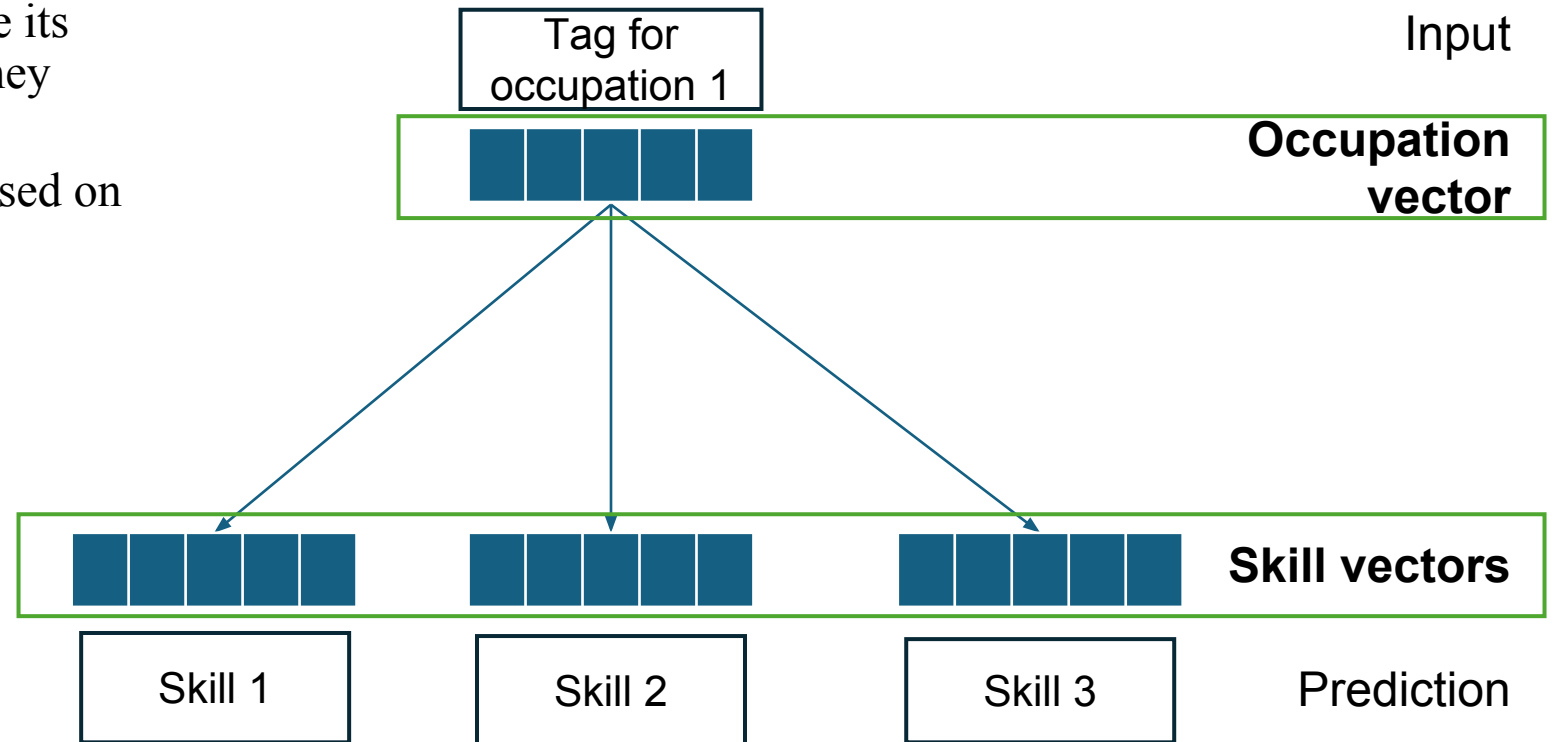
- > We use the simplest embedding model. It is a shallow neural network structure with one hidden layer: the Word2Vec model and its variant, the Doc2Vec model.
- > This is the basic and standard representation of the Word2Vec network. We use a variant of this model since our goal is to derive both skills and occupation embeddings.



## 2. Creation of a vector-based text representation of occupations and skills

### 1. Defining the model for creating vector-based text representations: Doc2Vec

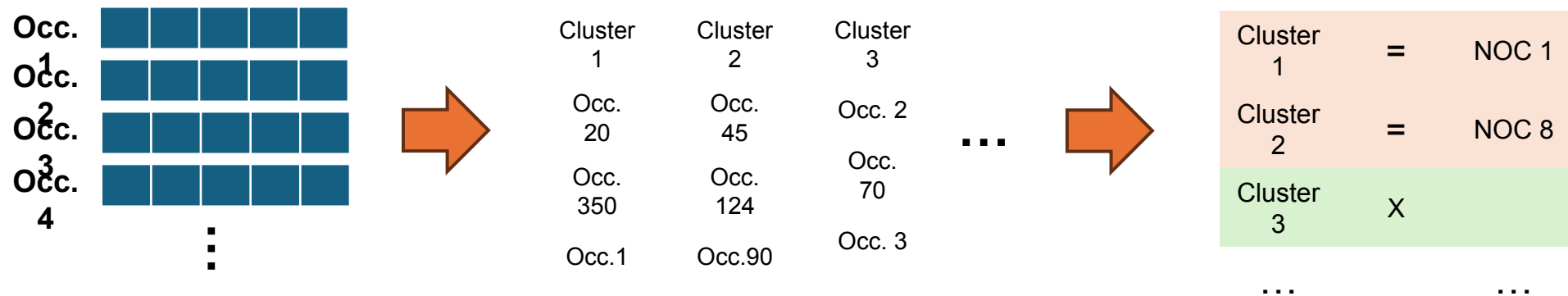
- > Version of the Doc2Vec model (PV-DBOW) in which a neural network is tasked to **predict skills randomly sampled** from an occupation document (list of skills) using as **input the occupation tag**.
- > Occupation and skill vectors are initially randomly assigned.
- > After several iterations:
  - **Occupation vectors** capture its meaning based on the skills they contain and their interaction.
  - **Skill vectors** are adjusted based on their appearance across
- > The model does not consider the order in which words appear in a document.



## 2. Creation of a vector-based text representation of occupations and skills

### 2. Creating a customised routine for selecting the best model hyperparameters

- > The D2V model requires defining some hyperparameters before running. The selection of these parameters affects both the performance of the model and the accuracy of the resulting vectors.
- > There is no standard benchmark for selecting optimal parameters:
  - > We developed a **routine that clusters the occupation vectors and compares the results with the NOC broad occupational categories** (1-digit) to select the best model.



- > Considerations:
  - > **Exact match** between clusters of occupations and NOC categories should not be expected as the latter consider other characteristics (i.e. occupations in NOC categories not always share the same skills, but also other work-related traits).
  - > This exercise, however, assumes that skill demands are a reliable predictor of most part of NOC categories.
  - > We expect that the accuracy of the information contained in vectors will improve as data-driven occupational clusters align more with NOC.

## 2. Creation of a vector-based text representation of occupations and skills

We apply the following steps:

**A. Cluster occupation-level vectors derived from the model**

- > We clustered 511 occupation vectors into 9 clusters using a K-means algorithm. This is an unsupervised learning model that classifies a group of data points based on the distance of each point to clusters' centroids (mean or median).

**A. Define a metric to compare actual classification with clustering results.**

- > To compare the performance of different models, we need an indicator that allows for easy comparison between them. We used two indicators to assess how well the vectors (and the clustering algorithm) aligns with NOC broad categories:

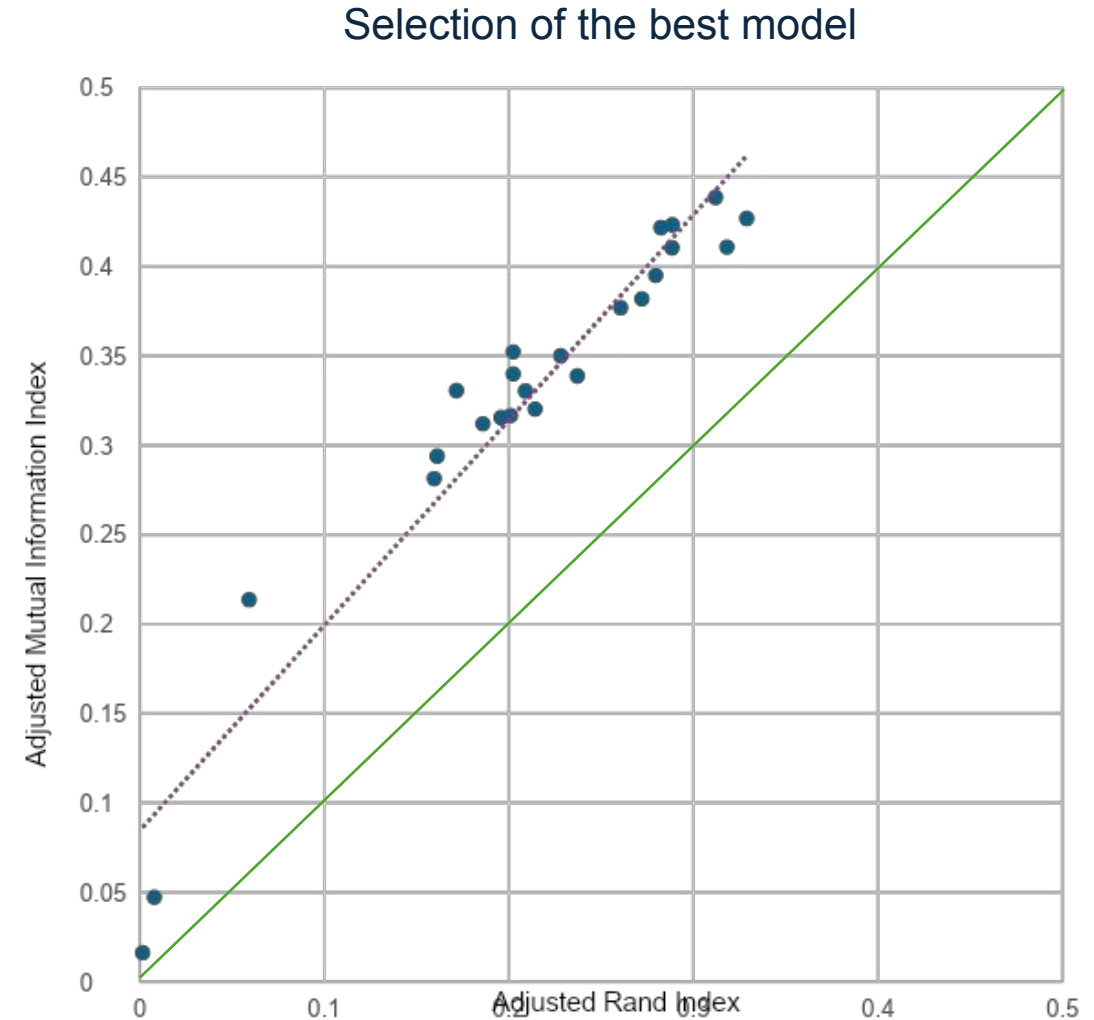
Adjusted Rand Index (ARI)	Adjusted Mutual Information Index (AMI)
Measures the similarity between two clusterings by considering all pairs of samples. It counts pairs assigned in the same or different clusters in the predicted and true clusterings, while accounting for random chance.  0 <input type="checkbox"/> Random or no agreement 1 <input type="checkbox"/> Perfect agreement.	Measures the agreement between two clusterings by calculating the mutual information between them. In other words, how much knowing the cluster assignment in one clustering helps to predict clusters in the other, accounting for random chance.  0 <input type="checkbox"/> No mutual information beyond random chance, 1 <input type="checkbox"/> Perfect agreement.

- > ARI is focused on the alignment at the NOC unit group level, while AMI has a more macro-view that gives more importance to NOC broad category alignment.

## 2. Creation of a vector-based text representation of occupations and skills

### C. Use the best model to generate occupation and skill vector dataset

- > The procedure iterates across multiple combinations of parameters. For each model, it trains the vectors, runs the clustering procedure and get the ARI and AMI metrics.
- > The first model evaluated is assumed to be the best, and it is replaced only if another model improves either of the two metrics.
- > We applied the procedure in a grid of parameters that produced 24 possible combinations (see figure).
  - > Parameters include vector length, common words undersampling, removal of low-frequency words, etc.
  - > The best models got an AMI around 0.4-0.45 and an ARI between 0.3-0.35. Both metrics are well correlated.
- > The results show that occupational vectors contain valuable information for replicating some NOC broad categories.



## 2. Creation of a vector-based text representation of occupations and skills

### D. Exploring the results

- > The vectors derived from the best model and the clustering procedure were able to replicate the occupations grouping of 4 NOC broad categories:
  - > Business and finance,
  - > Health,
  - > Natural resources and agriculture,
  - > Manufacturing and utilities.
- > The algorithm also accurately classifies some categories in two or three clusters:
  - > Natural and applied sciences,
  - > Trades, transport and equipment operators,
  - > Sales and services (see annex for more detail),
- > In some cases, it does not perform as well as in previous examples:
  - > Occupations in art and culture,
  - > Skilled trades cluster is too broad

Actual NOC broad occupation categories and predicted clusters.

		Predicted cluster								
		Recreation and Service Workers	Health and Social Care Support	Construction and Artistic Trades	Education and Teaching Professionals	Hospitality and Food Service Workers	Administrative and Financial Management	Logistics, Transport and equipment operators	Skilled Trades and Supervisors	IT and Engineering Professionals
Actual category	Business, finance and administration occupations	1	2		2		51	3		1
	Natural and applied sciences and related occupations						3	2	23	38
	Health occupations		43							
	Occupations in education, law and social, community and government services	2	13		6		15		16	
	Occupations in art, culture, recreation and sport	9	1	1	10		1		1	14
	Sales and service occupations	7	3		1	21	21	6	1	1
	Trades, transport and equipment operators and related occupations	1	1	29		6	5	37	13	2
	Natural resources, agriculture and related production occupations	1	1	2				7	17	
	Occupations in manufacturing and utilities	1		7		1	1	53	3	2

Note: Clustering aligns better with actual categories if most of the unit groups are aggregated in one single category. Since the NOC category 0 “Legislative and senior management occupations” has only 5 NOC unit groups (5-digit level), we did not consider this category as the K-means clustering algorithm can be sensible to large differences in clusters’ size. Cluster names were derived from asking an LLM to provide a name based on the occupations within each cluster.

Source: OECD calculations based on Lightcast data for Canada, 2023, and NOC taxonomy.

### 3. Producing granular skills profiles based on skill similarity scores

#### 1. Skill similarity as a proxy of skill relevance for occupations

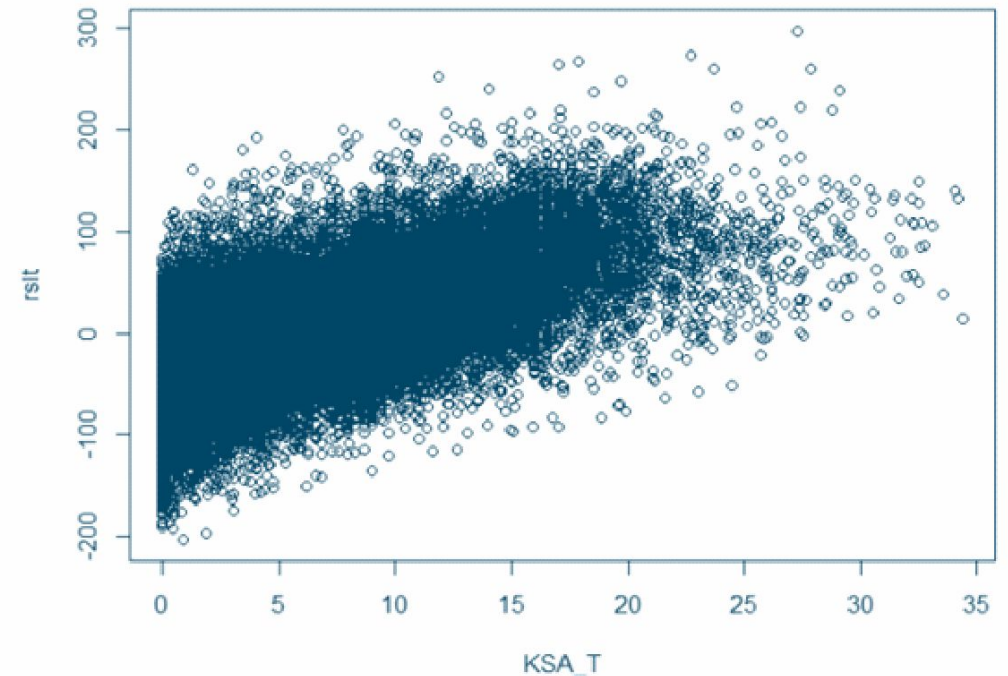
- Occupation and skill vectors retain semantic connections among them. Therefore, algebraic operations are expected return meaningful results.
- For instance,  
$$\text{vec}(\text{data}) + \text{vec}(\text{python}) \approx \text{vec}(\text{machine learning})$$
- Based on this property, we calculated a measure of similarity among vectors based on the cosine between any pair (A, B).
- The closer to 1, the more similar the pair of vectors.

>

$$\text{similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

- Empirical support to the hypothesis for which the skills similarity scores can indeed be used as a proxy of skill relevance
- We created two matrices summarising the results:
  - > **Semantic Skill Bundle Matrix (SSBM)**: Provides the relevance score for any occupation-skill pair in the model.
  - > **Occupation Similarity Matrix (OSM)**: Provides a measure of similarity between pairs of occupations.

Global correlation between SSBM relevance scores and O\*NET ranked values of importance and level



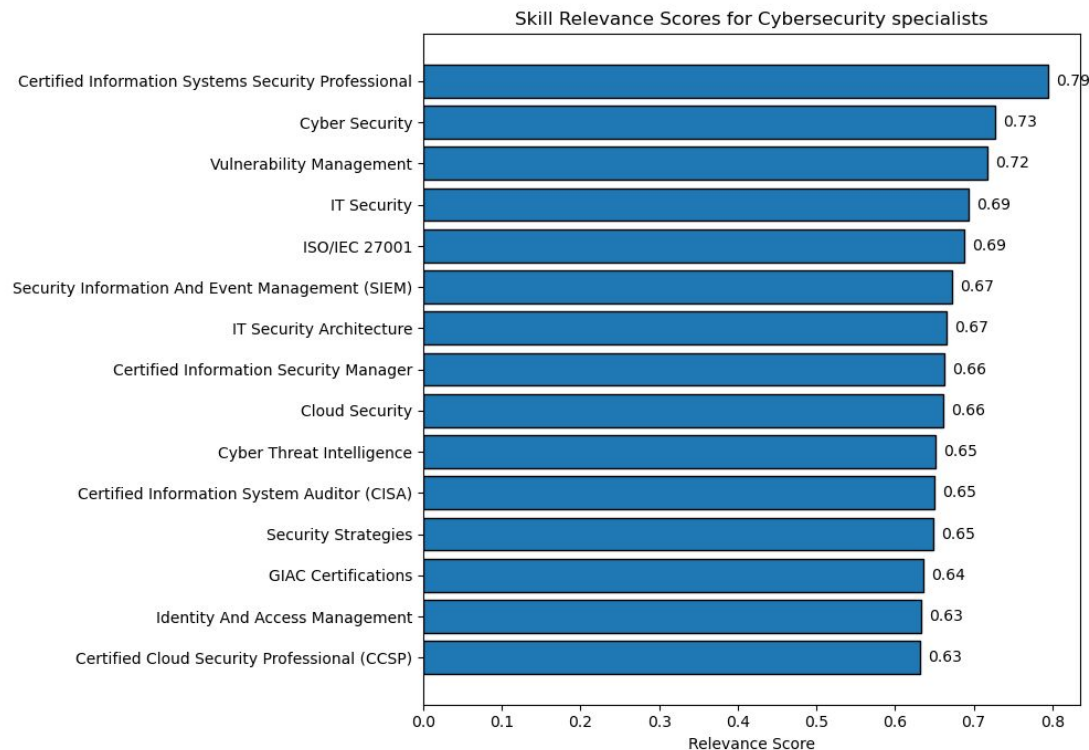
Note: Dots represent occupations at the 6th digit US SOC level. Each dot is the combination of two values: on the horizontal axis (KSA\_T) representing the O\*NET scores (importance\*level); on the vertical axis (rsit) the corresponding SSBM cosine similarity for every occupation.

Source: OECD calculations based on Lightcast data for the United States for the year 2019.

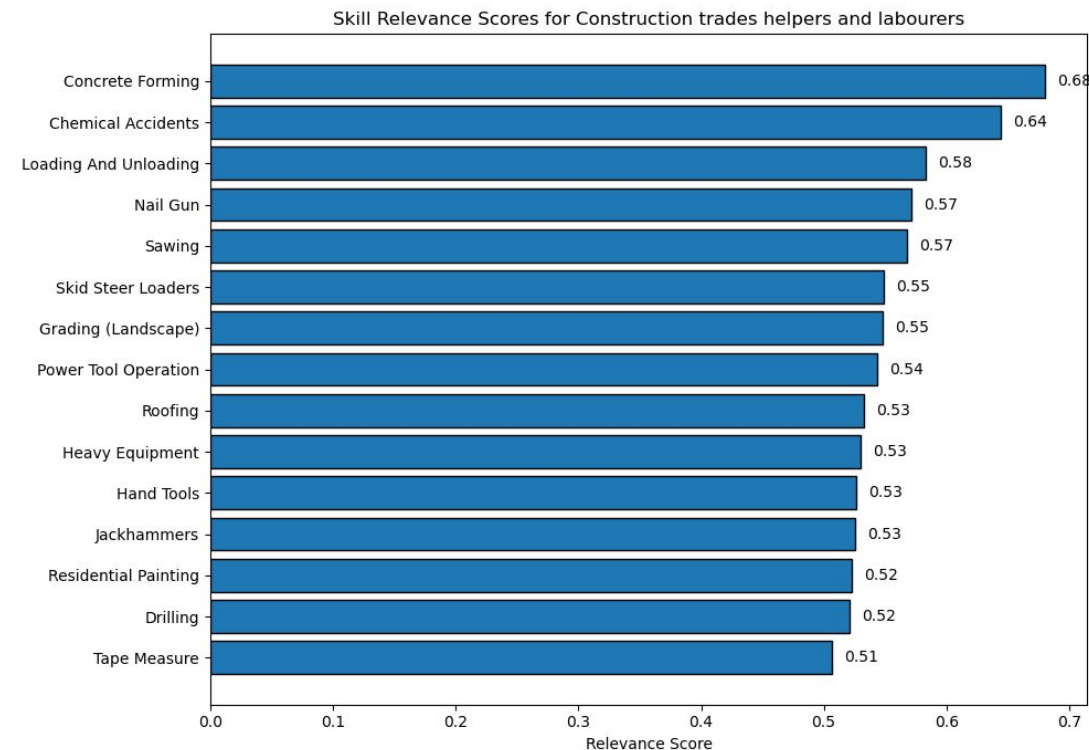
### 3. Producing granular skills profiles based on skill relevance scores

- Using the SSBM, it is possible to describe occupations based on their most relevant skills.
- Skills with very high relevance scores are typically quite specialised and relevant in a narrow set of occupations. Skills with lower similarity scores, instead, are likely to be found as relatively relevant in a larger set of occupations.

#### Cyber security specialists' skill profile

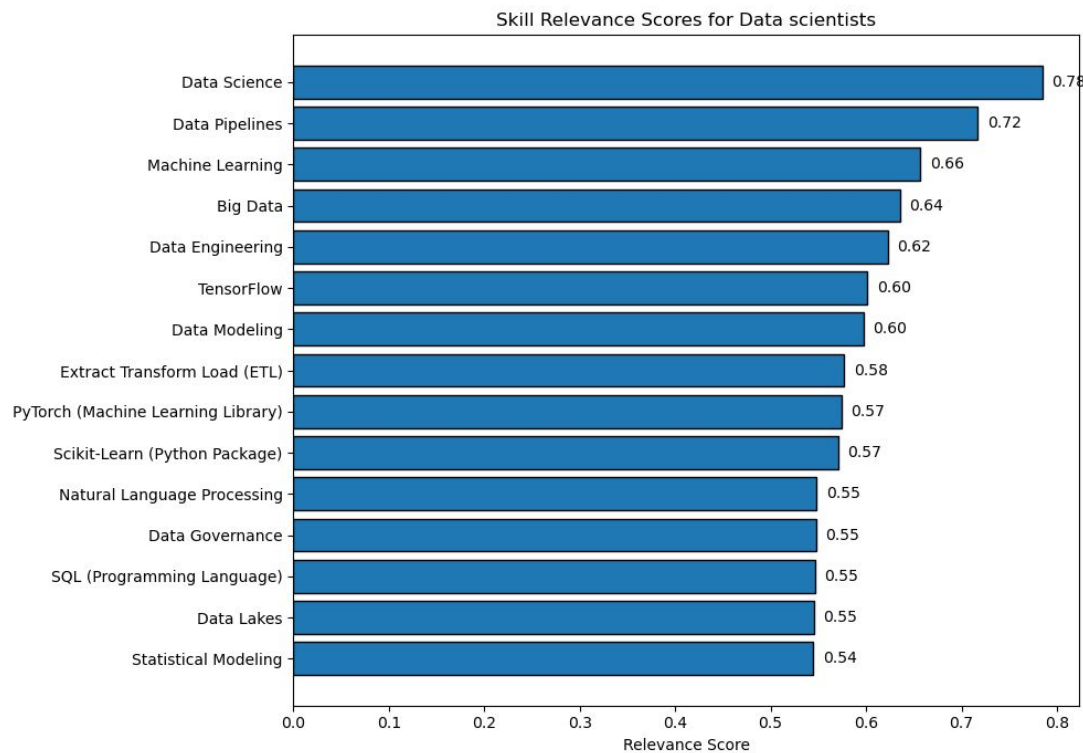


#### Construction trades helpers and labourers' skill profile

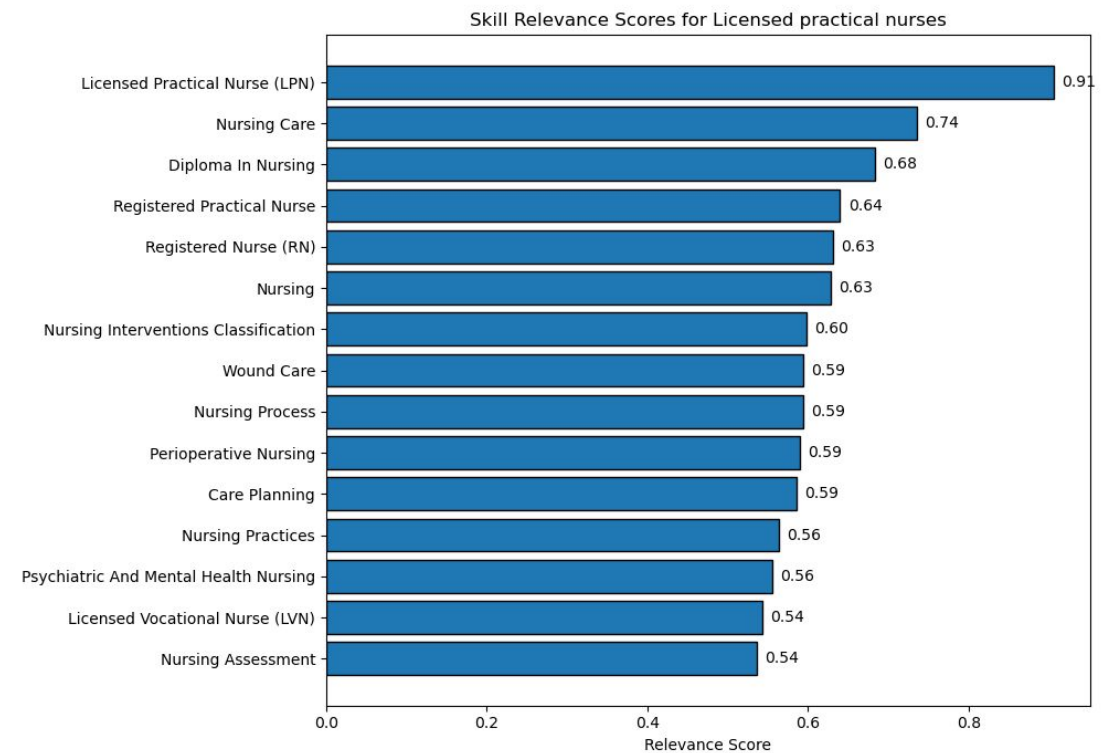


### 3. Producing granular skills profiles based on skill similarity scores

#### Data scientists' skill profile



#### Licensed practical nurses' skill profile





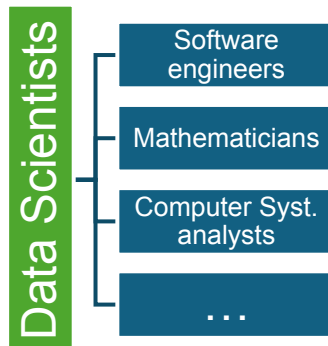
## Creation of retraining pathways

# Phase I, Output III - Creation of retraining pathways

- Occupation similarity and skill relevance scores shed light on feasible job transitions between occupations.
- We aim to compare pairs of occupations (origin and destination) based on their underlying skill profiles.

## Occupational Similarity Matrix

- > The OSM offers valuable information for identifying similarity across occupations in terms of global skills demand\*.
- > For each destination occupation, we consider the **10 most similar occupations with a similarity score higher than 0.25 (origin occupations)**.



## Skill relevance difference for measuring skill gaps

- > We then select the **15 most relevant specialised\*\* skills for the destination occupation** and calculate the difference in relevance scores ( $rel\_sc$ ) between the destination and origin occupations,

$$Diff_{i,j}^s = rel\_sc_i^s - rel\_sc_j^s$$
$$0.4 = 0.7 - 0.3$$

- > For the skill  $s$ , the destination occupation  $i$  and the origin occupation  $j$ .
- > The SSBM relevance scores are normalised to produce meaningful comparisons, with values ranging from -1 to 1.
- > The differences in skill relevance approximate the size of the skill gaps between the two occupations → **how much more or less important a particular skill in the destination occupation is compared to the origin occupation.**

## Training efforts

- > Estimated skill gaps allow for defining, **at the skill level, areas of training effort** for someone in the origin occupation to transition to the destination occupation:

Range	Indicator
Below 0.2	Low training effort
Between 0.2 and 0.5	Medium training effort
Higher than 0.5	High training effort

- > Results are presented using these indicators in a heatmap.

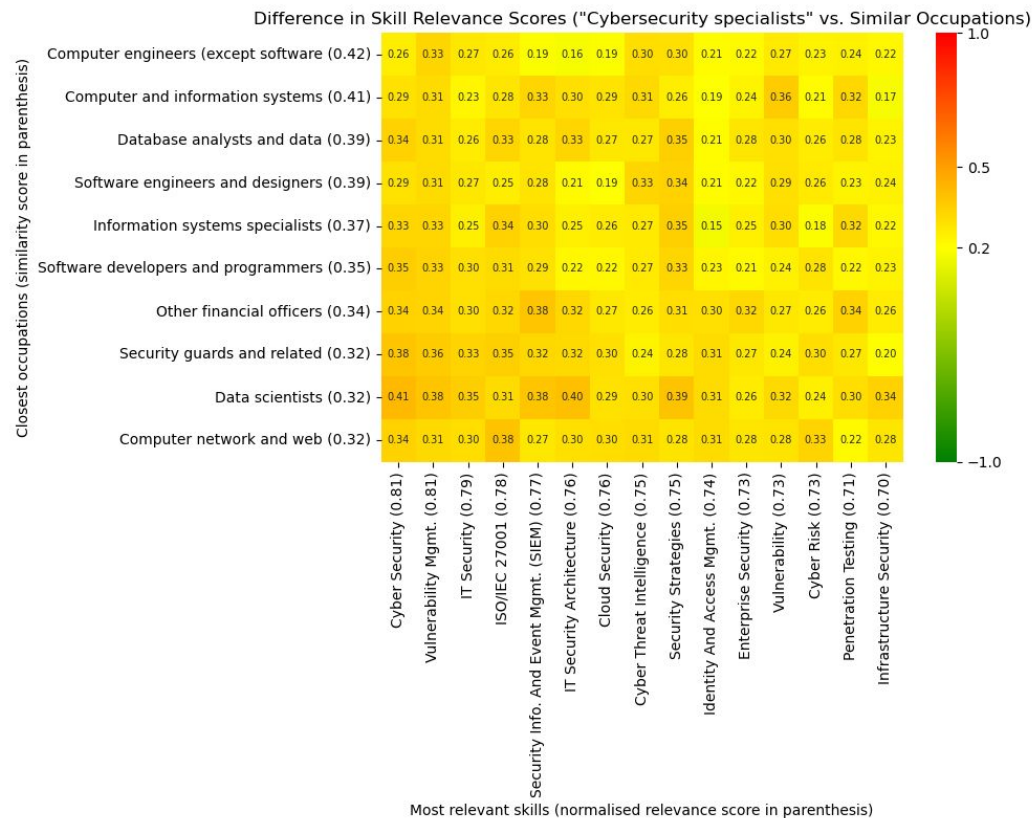
\* Previous work from the OECD has used alike methodologies (i.e. Euclidean distance) to estimate similarity in skill requirements between occupations (Tuccio et al. (2023) and OECD Employment Outlook (2024)).

\*\* Skills are divided into three categories: Specialised, Common and Certifications.

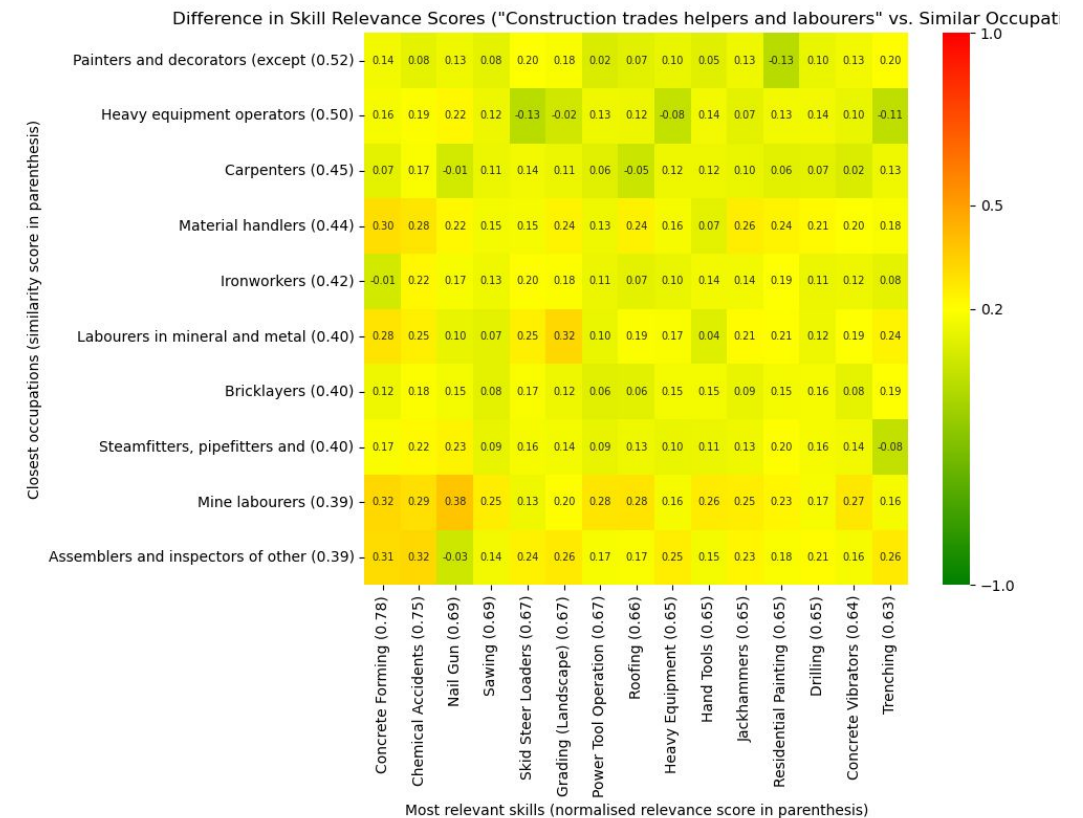
# 1. Occupation similarity, skill relevance scores and feasible job transitions

- The heatmaps below summarise the retraining pathways for a selected destination occupation.
- Results suggest that the training effort to become cyber security specialists (starting from occupations such as computer engineers etc.) is typically medium-high (0.3 - 0.4). To become a constructions helper, instead, the training needed is typically lower (range below 0.2).

## Destination occupation: Cyber security specialists



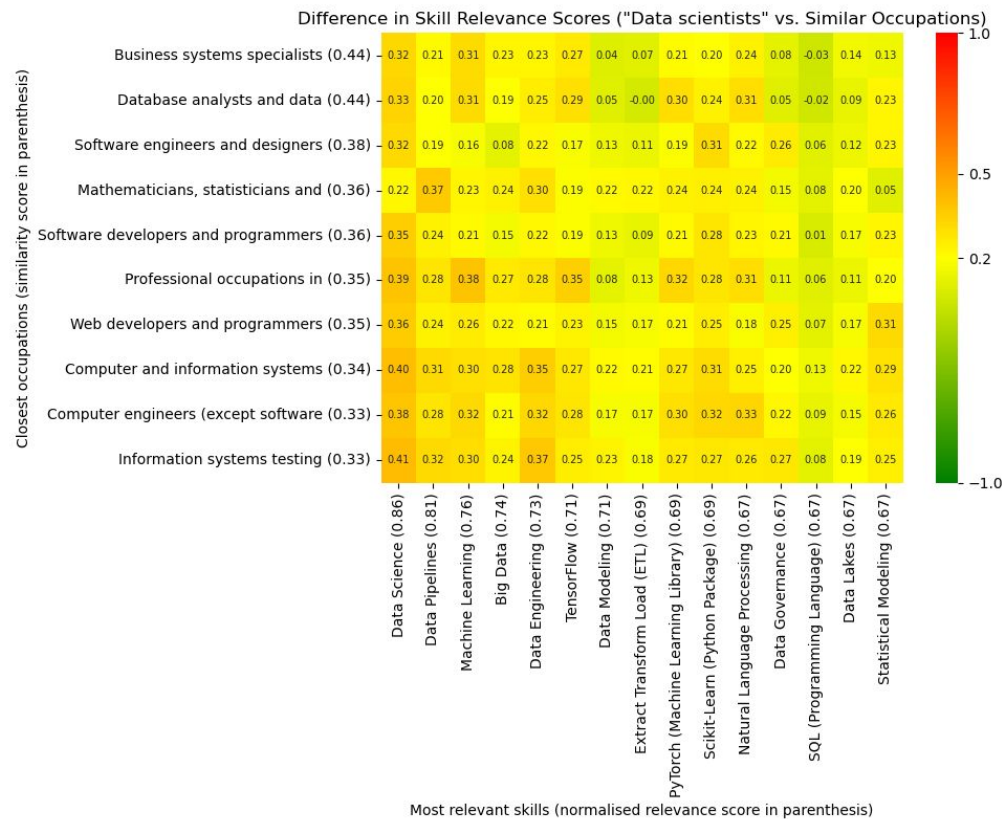
## Construction trades helpers and labourers



# 1. Occupation similarity, skill relevance scores and feasible job transitions

- This also speaks to the specificity of each occupation, as more specialised occupations are more difficult to transition into.
- The results suggest that data scientists are more specialized than licensed practical nurses, as the training required to transition from an origin occupation to becoming a data scientist is greater.

## Destination occupation: Data scientists



## Licensed practical nurses

